Association Mapping Analysis of a Core Collection of Flax (*Linum usitatissimum* L.)

BY

BRAULIO JORGE SOTO-CERDA

A Thesis
Submitted to the Faculty of Graduate Studies of the
University of Manitoba
in Partial Fulfilment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

Department of Plant Science
University of Manitoba
Winnipeg, Manitoba, Canada

THE UNIVERSITY OF MANITOBA
FACULTY OF GRADUATE STUDIES
COPYRIGHT PERMISSION

Association Mapping Analysis of a Core Collection of Flax (*Linum usitatissimum* L.)

BY

BRAULIO JORGE SOTO-CERDA

A thesis/Practicum submitted to the Faculty of Graduate Studies

of The University of Manitoba

in partial fulfilment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Acknowledgements

This dissertation was possible because God guided every step of my way, especially since I arrived to Canada. Thanks God for your blessings.

> "*Venid a mí todos los que estáis trabajados y cargados, y yo os haré descansar. Llevad mi yugo sobre vosotros, y aprended de mí, que soy manso y humilde de corazón; y hallaréis descanso para vuestras almas; porque mi yugo es fácil, y ligera mi carga*." (Mateo 11:28-30).

First and foremost, my sincere acknowledgements to my major supervisor, Dr. Sylvie Cloutier, for the opportunity to participate in the TUFGEN project, to develop my research in an exciting topic and to interact with scientists involved in flax breeding and research. Her priceless suggestions and support had an immense impact on my scientific formation, my future goals and on the success of this thesis.

I would like to thank the Bergen family for its warm reception and support at the beginning of my stay in Canada.

My sincere gratefulness to my committee members: Dr. Dana Schroeder, Dr. Peter McVetty and Dr. Curt McCartney for their guidance and comments to improve my research.

This work could not have been completed on time without the help of the Cloutier's lab members. My thanks and recognition to Natasa Radovanovic, Elsa Reimer, Mitali Banik, Kerry Ward, Evelyn Miranda, Andrzej Walichnowski, Santosh Kumar, Frank You and summer students for their technical and intellectual support. My special thanks to Dr. Raja Ragupathy, for being a role model, for his sincere friendship and advices during these years we shared office. I also thank the flax breeders Drs. Gordon

Rowland, Scott Duguid and Helen Booker, and their breeding teams at the Morden Research Station and the Crop Development Centre for technical assistance.

I wish to extend my sincere appreciation to the commissionaires at the Cereal Research Centre Ken Miller, Joseph Lysy and Robert Erstelle, for their generous friendship.

I acknowledge Becas Chile – Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) for the Financial support. I also acknowledge the Agriaquaculture Nutritional Genomic Center (CGNA), especially its director Dr. Haroldo Salvo-Garrido for his unconditional support since we met in 2002.

My special thanks to Sandra Quiroga who supported me during my stay in Canada with her prayers and love. This achievement is as mine as hers.

Finally my thanks to my mother and grandmother, who have been patiently waiting for me in Chile, wishing me the best in my studies.

BRAULIO JORGE SOTO-CERDA

Dedicated to

“God”

“My family”

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

AEC	Average Environment Coordinate
AM	Association Mapping
AFLP	Amplified Fragment Length Polymorphism
ANOVA	Analysis of Variance
AMMI	Additive Main effect and Multiplicative Interaction
ASV	AMMI's Stability Value
BPA	Bolls Per Area
CC	Case-Control
CG	Candidate Gene
DGAT	Acyl-CoA:Diacylglycerol Acyltransferase
DH	Doubled Haploid
EMS	Ethyl Methanesulphonate
EST	Expressed Sequence Tag
EW	Ewens-Watterson
FA	Fatty Acid
FAD	Fatty Acid Desaturase
FDR	False Discovery Rate
FL	Flowering
GC	Genomic Control
GE	Genotype x Environment interaction
GLM	General Linear Model
GO	Gene Ontology

GWAS	Genome Wide Association Study
IBD	Identical By Descent
IBS	Identical By State
IOD	Iodine Value
IPCA	Interaction Principal Component Analysis
IRAP	Inter Retrotransposon Amplified Polymorphism
ISSR	Inter Simple Sequence Repeat
LD	Linkage Disequilibrium
LDG	Lodging
LG	Linkage Group
LIN	Linolenic acid
LIO	Linoleic acid
MAD	Modified Augmented Design
MAF	Minor Allele Frequency
MAGIC	Multiparent Advanced Generation Intercross
MAS	Marker Assisted Selection
MB	Manitoba
MLM	Mixed Linear Model
NAM	Nested Association Mapping
NCBI	National Center for Biotechnology Information
NJ	Neighbor Joining
OIL	Oil content
OLE	Oleic acid

PAL	Palmitic acid
PB	Plant Branching
PCA	Principal Component Analysis
PCoA	Principal Coordinate Analysis
PGRC	Plant Gene Resources of Canada
PH	Plant Height
PIC	Polymorphism Information Content
QQE	QTL main effect and QTL by Environment interaction
QTL	Quantitative Trait Loci
RAPD	Random Amplified Polymorphic DNA
RFLP	Restriction Fragment Length Polymorphism
RIL	Recombinant Inbred Line
SA	Structured Association
SDG	Secoisolariciresinol Diglucoside
SK	Saskatchewan
SNP	Single Nucleotide Polymorphism
SPB	Seeds Per Boll
SSIPCA	Sum of Square Interaction Principal Component Analysis
SSR	Simple Sequence Repeat
STE	Stearic acid
TDT	Transmission Disequilibrium Test
TSW	Thousand Seed Weight
TUFGEN	Total Utilization Flax GENomics

WGS	Whole Genome Shotgun

LIST OF APPENDICES

Appendix Page

ABSTRACT

Linseed oil (*Linum usitatissimum* L.) is valued for its food and non-food applications. Although Canada is the world's largest linseed producer and exporter, linseed remains a minor crop in part because its yield has been stagnating over the last decade compared to other oilseeds. Narrow genetic base, absence of an efficient hybrid production system and limited genomic tools for linseed breeding are the main factors hindering yield and quality improvements. Here, we characterized the Canadian flax core collection of 407 accessions with 448 genome-wide simple sequence repeat markers and, using association mapping (AM), we demonstrated its potential for the improvement of seed quality and agronomic traits.

Genetic structure analyses assigned all accessions to two major groups that were weakly differentiated (F_{ST} = 0.094). Genetic diversity was abundant in the total panel (5.32 alleles per locus) with weak familial relatedness (mean = 0.287) for most individual pairs. Linkage disequilibrium (LD) decayed relatively quickly with an average genome-wide LD of ~1 cM.

AM for seven seed quality traits including oil content (OIL), palmitic acid (PAL), stearic acid (STE), oleic acid (OLE), linoleic acid (LIO), linolenic acid (LIN) and iodine value (IOD) identified nine stable candidate QTL. LIO and LIN QTL co-localized with previously identified QTL and some mapped in the vicinity of genes known to be involved in the fatty acid biosynthesis pathway.

AM conducted for nine agronomic traits including yield, bolls per area (BPA), seeds per boll (SPB), thousand seed weight (TSW), start of flowering (FL5%), end of

flowering (FL95%), plant height (PH), plant branching (PB) and lodging (LDG) identified twelve significant marker-trait associations for six of the traits. The associated markers explained between 0.5 to 18.5% of the phenotypic variation, with Lu526 and Lu2532 associated with TSW and Lu943 associated with flowering being the most promising for marker-assisted selection. Statistical simulation for five markers associated with TSW indicated that the favorable alleles have additive effects. None of the accessions carried the five favorable alleles but a few breeding lines had four, indicating that further improvement of TSW and yield could be achieved through marker assisted breeding.

FOREWORD

The thesis follows the paper style format recommended by the Plant Science Department and the Faculty of Graduate Studies of the University of Manitoba. The thesis has seven chapters: a general introduction, a literature review, three manuscripts, a general discussion and conclusion and the literature cited. Manuscripts were formatted following the guidelines of Theoretical and Applied Genetics. Each manuscript contains abstract, introduction, material and methods, results, discussion and conclusion or final remarks.

1.0 GENERAL INTRODUCTION

1.1 Flax

Flax (*Linum usitatissimum* L.) is a diploid (2n=2x=30) self-pollinated species with a genome size of ~370 Mb (Ragupathy et al. 2011). The species is believed to have originated in either the Middle East or Indian regions (Vavilov 1951). Morphological, cytological and molecular evidence suggest that the wild progenitor of cultivated flax is pale flax (*L. usitatissimum* L. subsp. *angustifolium* (Huds.) Thell.) which are interfertile (Gill and Yermanos 1967; Diederichsen and Hammer 1995; Fu et al. 2002).

Flax is one of the first crops domesticated by humans as early as 8,000 years ago, being a source of oil and fibre to early civilizations (van Zeist and Bakker-Heeres 1975). The two main morphotypes of cultivated flax are linseed (oil morphotype) and fibre flax (fibre morphotype). Linseed plants are shorter, more branched, larger seeded and grown over a wider area in continental climate regions such as Canada, India, China, the United States and Argentina (Green et al. 2008). The fibre morphotype, taller and less branched, is grown in the cool-temperate regions of China, the Russian Federation and Western Europe (Green et al. 2008).

Linseed produces valuable and unique oil with industrial, food and nutraceutical end-uses and is also a potential source of dietary fibre. A unique feature of linseed is the possibility of whole plant exploitation with minimal waste products (Czemplik et al. 2011). Its high nutritional value makes it ideal for the food, pharmaceutical and health care industries. Linseed is a rich source of soluble fibre and antioxidants (Westcott and Muir 2003), and its oil has a high concentration of omega-3 alpha linolenic acid

recognized for its health benefits (Przybylski 2001). Linseed stems are used in paper, technical fibre derivatives and biofuel (Diederichsen and Ulrich 2009; Cullis 2011).

In 2011, the total world production of linseed reached ~1.6 million tonnes, with Canada (~23%), China (~21%) and The Russian Federation (~14%) being the main producers (FAOSTAT 2013). In the same year, of the 368,000 tonnes of linseed produced by Canada, approximately 265,000 were harvested in the western Canadian provinces of Saskatchewan (67%), Alberta (18%) and Manitoba (14%) (http://www.canadagrainscouncil.ca/).

1.2 Flax genetic diversity

Genetic diversity represents the evolutionary potential of crops to cope with environmental changes (Sundar 2011). Determination of genetic diversity in flax was first performed using morphological parameters (Diederichsen 2001) and isozyme markers (Månsby et al. 2000). The use of DNA-based markers to study flax diversity was first reported by Oh et al. (2000). A few studies have attempted to reveal genetic relationships, extent of variation and genetic erosion across different flax collections based on random amplified polymorphic DNA (RAPD) markers (Fu et al. 2002, 2003; Diederichsen and Fu 2006). Amplified fragment length polymorphism (AFLP) and inter-simple sequence repeat (ISSR) markers have also been utilized (Spielmeyer et al. 1998; Everaert et al. 2001; Uysal et al. 2010). The ISSR technique was first applied in flax genetic studies by Wiesnerová and Wiesner (2004). Rajwade et al. (2010) later applied it to study the genetic diversity among Indian flax accessions. Other studies have used complementary approaches such as cytogenetic markers to reveal genetic variation

among flax genotypes and other species of the genus *Linum* (Muravenko et al. 2010; Rachinskaya et al. 2011).

Simple sequence repeat (SSR) markers consist of short tandem repeats characterized by high polymorphism, reproducibility, reliability, mostly co-dominant inheritance and relatively wide genome distribution (Powell et al. 1996) making them suitable for genetic diversity assessment in plants. The insufficient number of SSRs in flax has hindered the measurement of its genetic variation on a genomic scale. Significant numbers of SSRs in flax have only been developed in the last few years (Wiesner et al. 2001; Roose-Amsaleg et al. 2006; Cloutier et al. 2009, 2012a; Deng et al. 2010, 2011; Bickel et al. 2011; Soto-Cerda et al. 2011a, b; Kale et al. 2012). Flax genetic studies based on SSRs derived from expressed sequence tags (EST-SSRs) have been carried out by Cloutier et al. (2009) and Fu (2011) among a number of types including Canadian cultivars, dehiscent, pale, oil and fibre flax. Smỳkal et al. (2011) used retrotransposon-based markers to study the genetic diversity among 708 flax accessions. Taken together, these studies showed that most of the flax germplasm evaluated around the world has narrow genetic diversity. Several independent reports reiterated the narrow genetic base of Canadian flax cultivars (Fu et al. 2002, 2003; Cloutier et al. 2009) which is an impediment to further breeding progress.

In Canada, a world collection of approximately 3,500 accessions of flax traditionally deployed in flax breeding through a variety of conventional strategies, is maintained by Plant Gene Resources of Canada (Diederichsen 2007). The phenotypic and molecular characterization of the world collection enabled the selection of the Canadian flax core collection, a resource that will facilitate access to the diversity harboured in the

whole collection with the ultimate goal of flax genetic improvement (Diederichsen et al. 2013).

1.3 Flax genetic improvement

Conventional breeding techniques have been employed in flax for over a century and have been successful in developing new cultivars with durable resistance to the major wilt and rust diseases, improved lodging resistance, adapted crop phenology to match regional growing seasons and greater yield stability (Green et al. 2008). Flax breeding procedures are usually divided into three general categories: introduction, selection and hybridization followed by selection. Selection from introduced accessions and pedigree selection following hybridization have been the predominant methods of flax breeding (Duguid 2009). Prior to 1936, most of the cultivars recommended for production in North America were derived by mass or single plant selection from introduced accessions. Although many different breeding procedures and combinations thereof are used by flax breeders today, a universal feature of all current programs is hybridization followed by selection (Duguid 2009).

An understanding of the genetic bases of desirable traits is of practical value to the flax breeder because such information assists in the design of crosses and subsequent selection strategies. Pedigree selection, backcrossing, single seed descent, doubled haploid and mutation breeding methods have all been used in Canada to enhance breeding efficiency for agronomic, phenological and seed quality traits (Green et al. 2008; Duguid 2009).

1.3.1 Agronomic traits

High seed yield is an important objective for flax improvement. Over the past century, flax breeders have been successful in exploiting the germplasm base to produce the currently available cultivars (Duguid 2009). Although it is possible to breed for high yield per se, flax breeders also focus on yield components including number of bolls per plant, number of bolls per unit area, number of seeds per boll and thousand seed weight. The high yielding and broadly adapted cultivar CDC Bethune released in 1998 (Rowland et al. 2002), remains the single most important variety in terms of acreage in Canada, and it is still used as a check in the Canadian flax registration trials. This is symptomatic of the yield plateau observed over the last decade or two in flax which lags behind other oilseeds, particularly rapeseed (canola). During the last ten years, the overall flax yield has almost stagnated averaging 1.2 T/Ha (Statistics Canada; http://www.statcan.gc.ca).

Plant height is an important selection criterion in linseed breeding since shorter cultivars generally exhibit reduced risk of lodging and easier handling during harvest minimizing yield losses and seed quality reductions. In fibre flax, selection of tall plants is prioritized because the stem length largely determines the quality of the fibre for the textile industry (Diederichsen and Richards 2003).

Flax diseases are a potential constraint to production in nearly all areas of the world where flax is produced, and genetic resistance is the most cost effective method to prevent yield and quality losses. The fungal diseases flax rust (*Melampsora lini*), fusarium wilt (*Fusarium oxysporum* f.sp. *lini*), powdery mildew (*Oidium lini*) and pasmo (*Septoria linicola*) have received the most attention (Duguid 2009). The occurrence, severity and importance of flax diseases vary among flax-growing areas of the world

(Rashid 2003). Various sources of resistance or tolerance to these diseases have been identified and are being introgressed into elite cultivars (Duguid 2009).

1.3.2 Phenological traits

Flowering time and maturity, which are generally positively correlated, are two important traits for flax breeders because both are critical for the optimal production of seeds under specific environmental conditions (Jarillo and Piñeiro 2011). The breeding of early flowering and early maturity flax provide them with a better opportunity to escape damage from abiotic stresses such as drought, cold and frost that are detrimental to yield and quality (Duguid 2009).

1.3.3 Seed quality traits

Oil content is the most important seed quality trait for linseed and current Canadian linseed cultivars contain ~45-50% oil (Cloutier et al. 2011). Linseed oil is composed of five main fatty acids: palmitic (C16:0, ~6%), stearic (C18:0, ~2.5%), oleic (C18:1, ~19%), linoleic (C18:2, ~14%) and linolenic (C18:3, ~55%) (Westcott and Muir 2003), which largely define its nutritional quality and end-use functionality.

Linseed breeders have focused mainly on maintaining a high linolenic acid content, while it confers linseed oil its oxidative instability, it simultaneously determines to a large extent its primary utilization in the drying oil industries (Cullis 2007). Advantages of fatty acid modifications in linseed have been defined and research efforts are underway to develop germplasm with different fatty acid profiles for use by the chemical and food industries (Duguid 2009). High linolenic acid content (>65%)

germplasm is available (Friedt et al. 1995; Kenaschuk 2005) but agronomic improvement of many of these sources is needed to achieve adaptability. The first high linolenic acid linseed cultivar NuLin^{TM} 50 was registered in Canada by Viterra in 2008 (http://www.viterra.ca). Reduced linolenic acid content (2-4%) and high linoleic acid content (71%) cultivars, commonly termed solin to differentiate them from traditional linseed cultivars, have been developed by mutation breeding (Green 1986; Rowland 1991).

Oil content and fatty acid composition are highly dependent on the environments where cultivars are grown (Casa et al. 1999; Fofana et al. 2006). In Canada, oil content can vary up to 15% (range 35-50%) on individual farm samples (Duguid 2009) and the percentage of linolenic acid can be as much as 5% higher in cool environments (Fofana et al. 2006).

The crude protein content of linseed is relatively high (~23%) and consists of approximately 20% albumin and 80% legumin-like proteins (Oomah and Mazza 1993) with an amino acid profile similar to soybean but with relatively higher levels of aspartic acid, glutamic acid and arginine (Oomah 2001). The linseed meal protein content ranges between 43 and 46%, hence, linseed meal is an excellent source of proteins for livestock (Duguid 2009). Sources of higher protein content are available. For example, line FP2188 contains 47% meal protein and, having also high oil content, it does not follow the typical inverse relationship between oil and protein contents observed in other oilseeds (Duguid 2009). The relatively high protein content of linseed with other components has roused interest for its potential effect in reducing hypertension and heart diseases (Oomah 2001). Linseed mucilage is an excellent source of dietary fibre. Due to its specific biological

activities, mucilage can be used as a thickening agent (Diederichsen et al. 2006), as a substitute to chemical additives for food preservation (Guilloux et al. 2009) and as an excipient in drug formulations (Avachat et al. 2011). Linseed mucilage can vary between 3.6 and 8% of the seed weight (Oomah et al. 1995) and it exists in two water-soluble forms. The first form is composed by rhamnogalacturonan I, which is easily separated from the seed coat at room temperature (Naran et al. 2008). The second is composed by arabinoxylan that tightly adheres to special secretory cell walls, which can be efficiently extracted at higher temperature (100 ºC) (Naran et al. 2008; Barbary et al. 2009). Although the chemical properties of linseed mucilage have been studied (Bhatty 1993; Guilloux et al. 2009; Rasmussen and Meyer 2010), its genetic architecture is poorly understood. In *Arabidopsis*, several genes affecting the mucilage synthesis and secretion pathway have been characterized (Western et al. 2001; Dean et al. 2007; Arsovski et al. 2009) and some of their orthologs have been identified in flax ESTs indicating that the mucilage pathway observed in *Arabidopsis* is present in flax (Venglat et al. 2011).

The phytochemical lignans are abundant in linseed with secoisolariciresinol diglucoside (SDG) being the principal type recognised for its antioxidant and anticancer activities (Touré and Xueming 2010). Of the six major groupings of flax, the Indian, Mediterranean and spring collections contain accessions with the highest SDG levels, while winter, fibre and forage types contain on average less SDG. The SDG levels of Canadian cultivars range between 13 and 22 mg/g of seed (Duguid 2009).

1.4 Flax biotechnology and molecular breeding

Flax improvement through biotechnology has been based on mutation breeding, somaclonal variation, doubled haploid breeding and genetic transformation (Cullis 2011). Mutation breeding using ethyl methanesulphonate (EMS) has been successfully applied to alter fatty acid composition profiles (Green and Marshall 1984; Green 1986; Rowland and Bhatty 1990; Rowland 1991). Somaclonal variation induced through anther culture resulted in regenerants with valuable breeding features, including enhanced resistance to fusarium wilt (McHughen and Swartz 1984). Doubled haploid breeding was first described by Rajhathy (1976) in polyembrionic seeds of flax. More recently, haploid plants have been reported to be routinely produced by microspore culture providing an efficient system to produce homozygous progenies (Chen and Dribnenki 2004). Flax has been genetically transformed using both *Agrobacterium tumefaciens* and particle bombardment. The first trait engineered into flax was tolerance to the herbicide glyphosate (Jordan and McHughen 1988). The cultivar CDC Triffid, resistant to a sulfonylurea herbicide, was considered for commercial release in Canada in 1998, but it was not commercialized at the request of the flax industry because of the European Union's concern with genetically modified flax (Jhala et al. 2009). Efforts to enhance the thermoplastic properties of flax fibre were conducted by Wróbel et al. (2004) who expressed three bacterial genes encoding polyhydroxybutyrate (PHB), a nontoxic biodegradable thermoplastic agent with physical properties similar to the polymer polypropylene, but this effort did not translate in the release of a commercial variety.

Many important traits in flax such as yield, oil content, fatty acid composition, flowering time and some of the disease resistance mechanisms are under control of

quantitative trait loci (QTL). Breeding based only on conventional phenotypic selection is not efficient because quantitative traits are generally influenced by the environment, genotype by environment interactions and experimental errors, hampering genetic gain. A major breakthrough in the characterization of QTL was initiated by the development of DNA markers in the 1980s, and their use in the construction of linkage maps (Lander and Botstein 1989). In flax, only three individual linkage maps (Spielmeyer et al. 1998; Oh et al. 2000; Cloutier et al. 2011) have been published to date. The linkage map developed by Cloutier et al. (2011) had 113 markers mostly SSRs, while those of Spielmeyer et al. (1998) and Oh et al. (2000) were based on AFLP and RFLP/RAPD markers, respectively. The limitations of these maps reside in either or both the type and limited number of markers (Cloutier et al. 2012b) hindering the application of marker-assisted selection (MAS). Only two of these maps were used for QTL studies, positioning QTL for *Fusarium* wilt resistance (Spielmeyer et al. 1998) and QTL for iodine value, palmitic, linoleic and linolenic acids (Cloutier et al. 2011). The recent identification of a large number of polymorphic SSR markers (Cloutier et al. 2009, 2012a) has enabled the construction of a high density consensus linkage map of more than 700 SSR loci (Cloutier et al. 2012b) that will provide a reference for a wide variety of applications such as QTL mapping, association mapping, map-based gene cloning, phylogenetic analyses, anchoring of the whole genome shotgun sequence assembly and ultimately, MAS (Cloutier et al. 2012b).

1.5 The rationale and scope of the research

Linseed (*Linum usitatissimum* L.), with its high content of alpha linolenic acid, is a very unique oilseed crop. Although Canada is the world's largest linseed producer and

exporter (FAOSAT 2013), in Canadian agriculture, areas planted in linseed are less than wheat, barley, oats, canola and soybean (Statistics Canada; http://www.statcan.gc.ca). The limited genomic tools for linseed genetic applications and the narrow genetic base used for the development of Canadian linseed cultivars (Fu et al. 2002, 2003; Cloutier et al. 2009) have limited further yield and quality improvement reducing linseed competitiveness with other crops.

In 2009, the Total Utilization Flax GENomics (TUFGEN; http://www.tufgen.ca) project was initiated in Canada to generate genomics resources for flax and to apply them to an array of traits for the ultimate purpose of flax improvement. Among the genomics tools available for crop breeding, association mapping has emerged as a powerful option for the identification of QTL utilizing large populations with more allelic diversity than biparental populations. This improves the probability of QTL detection and provides the mapping resolution needed for effective MAS (Ersoz et al. 2009). Thus, the QTL identified with association mapping can be deployed in MAS, improving the efficiency of traditional linseed breeding and reducing the yield gap between linseed and other oil crops. Likewise, understanding the QTL architecture of seed quality traits could enable the modification of the fatty acid profiles to meet market needs in a timely manner.

The present research had the following objectives: (i) to evaluate the population structure and genetic diversity of the Canadian flax core collection, (ii) to determine the extent and variation of linkage disequilibrium in *L. usitatissimum*, (iii) to identify, through association mapping, genes/QTL related to agronomic and seed quality traits in *L. usitatissimum* and (iv) to validate the QTL-marker associations using linkage maps carrying mapped QTL related to agronomic and seed quality traits.

LITERATURE REVIEW

The literature review was published as a book chapter: Soto-Cerda BJ, Cloutier S (2012) Association mapping in plant genomes. In: Caliskan M (ed) Genetic diversity in plants. InTech, Rijeka, pp 29–54.

Braulio J. Soto-Cerda wrote the manuscript. Dr. Sylvie Cloutier supervised the work and co-wrote the manuscript.

2.0 LITERATURE REVIEW

2.1 Linkage mapping and association mapping

Linkage mapping has been a key tool for identifying the genetic basis of quantitative traits in plants. However, for linkage studies, suitable crosses, sometimes limited by low polymorphism or small population size, are required. In addition, only two alleles per locus and few recombination events are considered to estimate the genetic distances between marker loci and Quantitative Trait Loci (QTL), thereby limiting the mapping resolution. To circumvent these limitations, linkage disequilibrium (LD) mapping or association mapping (AM) has been used extensively to dissect human diseases (Slatkin 2008). AM has the potential to identify a single polymorphism within a gene that is responsible for phenotypic differences. AM involves searching for genotype-phenotype correlations among unrelated individuals. Its high resolution is accounted for by the historical recombination accumulated in germplasm collections. By exploiting broader genetic diversity, AM offers three main advantages over linkage mapping: mapping resolution, allele number and time saving in establishing a marker-trait association and its application in a breeding program (Flint-Garcia et al. 2003), hence there is tremendous interest in using AM to examine agricultural traits.

2.2 Linkage disequilibrium and association mapping concepts

The terms LD and AM have often been used interchangeably in the literature. However, they present subtle differences. According to Gupta et al. (2005), AM refers to the significant association of a marker locus with a phenotype trait while LD refers to the

non-random association between two markers (Fig. 2.1). Thus, AM is actually an application of LD. In other words, two markers in LD represent a non-random association between alleles, but do not necessarily correlate/associate with a particular phenotype, whereas association implies a statistical significance and refers to the covariance of a marker and a phenotype of interest. Likewise, the concepts of linkage and LD are commonly confused. Linkage refers to the correlated inheritance of loci through the physical connection on a chromosome, whereas LD refers to the correlation between alleles in a population (Flint-Garcia et al. 2003). Although tight linkage between alleles on the same chromosome generally translate into high LD, significant LD may also exist between distant loci, and even between loci located on different chromosomes. Both QTL and AM approaches are therefore based on LD between molecular markers and functional loci. In QTL mapping, LD is generated by the mating design while in AM, LD is a reflection of the germplasm collection under study (Stich and Melchinger 2010). In a mapping population, LD is influenced only by recombination in the absence of segregation distortion. In AM, LD may also be influenced by other forces such as selection, mutation, mating system, population structure, etc.

The concept of LD was first described by Jennings in 1917, and its quantification (D) was developed by Lewontin (1964). The simplified explanation of the commonly used LD measure, D or D' (standardized version of D), is the difference between the observed gametic frequencies of haplotypes and the expected gametic frequencies of haplotypes under linkage equilibrium.

$$D = P_{AB} - P_A P_B \qquad (1)$$

Where P_{AB} is the frequency of gametes carrying allele A and B at two loci; P_A and P_B are the product of the frequencies of the allele A and B, respectively. In the absence of other forces, recombination through random mating breaks down the LD with $D_t = D_0 (1 - r)^t$, where D_t is the remaining LD between two loci after t generations of random mating from the original D_0 (Zhu et al. 2008).

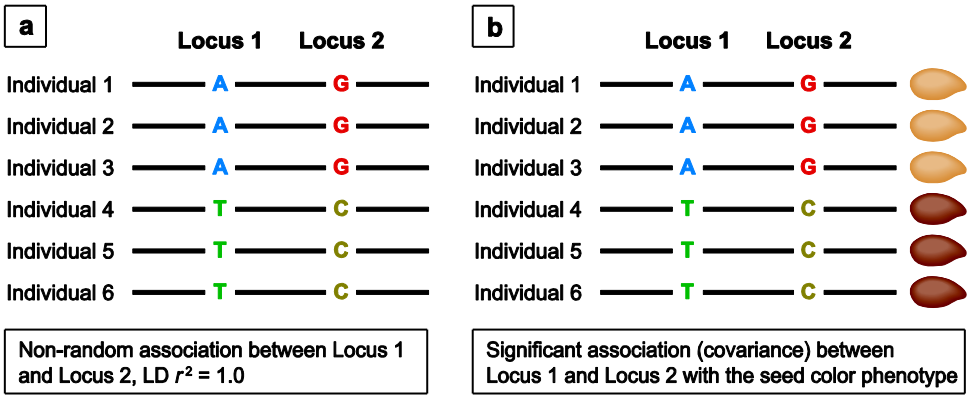


Fig. 2.1 Principles of linkage disequilibrium and association mapping. **a** Linkage disequilibrium. Locus 1 and Locus 2 present an unusual pattern of association between alleles A-G and T-C, which deviate from Hardy-Weinberg expectations, but without any statistical correlation with a phenotype. **b** Association mapping. Locus 1 and Locus 2 are in LD. Significant covariance with the seed color phenotype is considered evidence of association.

Several statistics have been proposed for LD, and these measurements generally differ in how they are affected by marginal allele frequencies and sample sizes. The two most utilized statistics for LD are D' (Lewontin 1964) and r^2, the square of the correlation coefficient between two loci (Hill and Robertson 1968), reflect different aspects of LD and perform differently under various conditions. D' only reflects the recombinational history and is therefore a more accurate statistic for estimating recombination differences, whereas r^2 summarizes both recombinational and mutational history (Flint-Garcia et al. 2003). For two biallelic loci, D' and r^2 have the following formula:

$$D' = |D| / D_{max} \quad (2)$$

$D_{max} = \min (P_A P_b, P_a P_B)$ if $D > 0$;

$D_{max} = \min (P_A P_B, P_a P_b)$ if $D < 0$

$$r^2 = D^2 / P_A P_a P_B P_b \quad (3)$$

D is limited because its range is determined by allele frequencies. *D′* was developed to partially normalize *D* with respect to the maximum value possible for the allele frequencies and give it a range between 0 and 1 (Zhu et al. 2008). The r^2 statistic has an expectation of $1/(1+4Nc)$, where *N* is the effective population size and *c* is the recombination rate, and it also varies between 0 and 1 (Hill and Robertson 1968). Choosing the appropriate LD statistics depends on the objective of the study. Most studies on LD in animal populations used *D′* to measure population-wide LD of microsatellite data (Du et al. 2007). However, *D′* is inflated by small sample size and low allele frequencies; therefore, intermediate values of *D′* are unsafe for comparative analyses of different studies and should be verified with r^2 before being used for quantification of the extent of LD (Oraguzie et al. 2007). Although r^2 is still considered to be allele frequency dependent, the bias due to allele frequency is considerably smaller than in *D′* (Ardlie et al. 2002). Currently, most LD mapping studies in plants use r^2 for LD quantification because it also provides information about the correlation between markers and QTL of interest (Flint-Garcia et al. 2003; Gupta et al. 2005). Typically, r^2 values of 0.1 or 0.2 are often considered the minimum thresholds for significant association between pairs of loci and to describe the maximum genetic or physical distance at which LD is significant (Zhu et al. 2008).

2.3 Visualization and statistical significance of LD

There are two common ways to visualize the extent of LD and the genomic regions or haplotype blocks found to be in significant LD. LD scatter plots are used to estimate the rate at which LD declines with genetic or physical distance (Fig. 2.2a). An average genome-wide decay of LD can be estimated by plotting LD values, from a data set covering an entire genome, against distance. These scatter plots are useful to determine the average effective distance threshold above which significant LD (commonly 0.5 for D' and 0.1 for r^2) is expected based on the curve of a nonlinear logarithmic trend drawn through the data points of the scatter plot (Breseghello and Sorrells 2006). Disequilibrium matrices or LD heat maps are also very useful for visualizing the linear arrangement of LD between polymorphic sites within a short physical distance such as a gene, along an entire chromosome or across the whole genome (Fig. 2.2b) (Flint-Garcia et al. 2003). LD heat maps are color-coded triangular plots where the diagonal represents ordered loci and the different intensity colored pixels depict significant pairwise LD level expressed as D' or r^2. Blocks of high intensity pixels afford an easy visualization of loci in significant LD. In Fig. 2.2b, the larger the blue blocks of haplotypes along the diagonal of the triangular plot, the higher the level and extent of LD between adjacent loci in the blocks, meaning that there has been either limited or no recombination since the LD block formation (Abdurakhmonov and Abdukarimov 2008). These graphical representations enable us to determine the optimum number of markers to detect significant marker-trait associations and the resolution at which a QTL can be mapped. Because LD estimation based on D' or r^2 can be sensitive to marker density, highly saturated and representative linkage groups are ideal for LD calculations.

The statistical significance of LD is typically determined using a χ^2 test of a 2 x 2 contingency table. A *p*-value threshold of 0.05 is often used to declare lack of independence of alleles at two loci, thus suggesting association (Gupta et al. 2005). From a 2 x 2 contingency table, the probability (*P*) of independence of alleles at two loci is generally calculated through a Fisher's exact test (Fisher 1935). Statistically significant LD can also be calculated using a multifactorial permutation analysis to compare sites with more than two alleles such as microsatellite markers. These statistical methods are implemented in software such as PowerMarker (Liu and Muse 2005) and TASSEL (Trait Analysis by aSSociation Evolution and Linkage) (Bradbury et al. 2007).

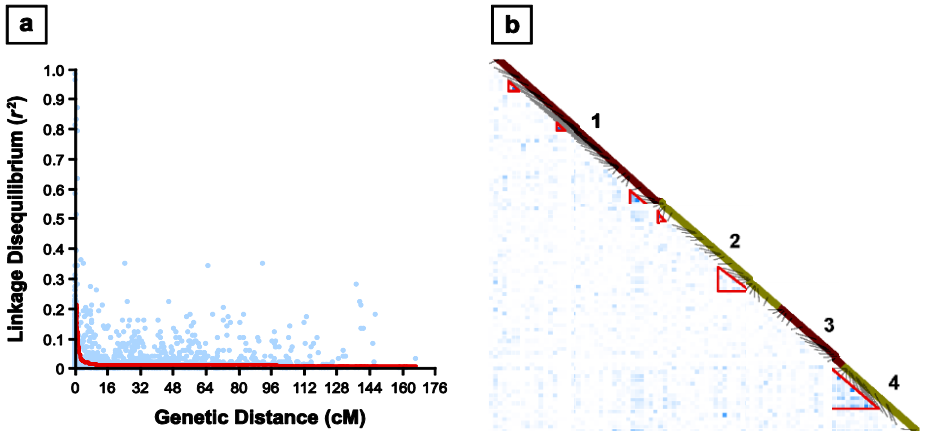


Fig. 2.2 Visualization of linkage disequilibrium in flax (*Linum usitatissimum* L.). **a** Scatter plot of LD decay (r^2) against genetic distance (cM), representing a measure of an average genome-wide LD. **b** Heatmap of LD variation between pairwise polymorphic loci of four linkage groups. Blocks in significant LD are highlighted by red triangles. LD distribution is heterogeneous within and between linkage groups.

2.4 LD variation as an effect of biological factors

2.4.1 Recombination

Several biological factors influence LD strength and its distribution across genomes. Many regions of the human genome display rates of recombination that differ significantly from the genome average recombination rate of 1 cM/Mb (Arnheim et al.

2003). These regions have been called “hotspots” and “coldspots” for high and low recombination rates, respectively. LD is strongly influenced by localized recombination rate and is correlated with other associated factors such as GC content and gene density (Dawson et al. 2002). In principle, local sequence features can affect LD directly and indirectly. For example, GC-rich sequences may be associated with higher rates of recombination and/or mutation, two phenomena that could directly lower surrounding levels of LD. Furthermore, in some protein-coding sequences, changes created by recombination or mutation may affect the fitness of an individual, and these sequences could be indirectly associated with unique patterns of LD as a consequence of natural selection (Smith et al. 2005).

Because LD is broken down by recombination, and recombination is not distributed homogeneously across the genome, blocks of LD are expected. Also, differences in LD between micro chromosomes and macro chromosomes have been reported (Stapley et al. 2010) as well as intra-chromosomal variation, where centromeric regions showed higher levels of LD. Teo et al. (2009) conducted a comprehensive analysis of genomic regions with different patterns of LD to unravel the consequences of this patterning for AM in human populations. Plant genomes have revealed similar general conclusions with regards to LD distribution. Inter-chromosomal LD variation has been reported in barley (*Hordeum vulgare*), maize (*Zea mays*), tomato (*Solanum lycopersicum*) and bread wheat (*Triticum aestivum*) (Malysheva-Otto et al. 2006; Yan et al. 2009; Zhang et al. 2010; Robbins et al. 2011), where it varied between less than 1 cM to more than 30 cM ($r^2 > 0.1$). As a consequence, investigation of LD variation at the

genome and chromosome scale to accurately estimate marker density for each chromosome is required to provide insights to the most cost-effective AM approach.

2.4.2 Mating system

The mating system has profound effects on LD (Myles et al. 2009). Selfing reduces opportunities for effective recombination because individuals are more likely to be homozygous than in outcrossing species (Flint-Garcia et al. 2003). In self-pollinated species such as rice (*Oryza sativa*), Arabidopsis (*Arabidopsis thaliana*) and wheat (*Triticum aestivum*) (Nordborg 2000; Garris et al. 2005; Zhang et al. 2010), LD extends much further as compared to outcrossing species such as maize (*Zea mays*) and rye (*Secale cereale*) (Tenaillon et al. 2001; Li et al. 2011c). As a result, genetic polymorphisms tend to remain correlated, and LD is expected to be maintained over long genetic or physical distances (Gaut and Long 2003). However, because LD declines more rapidly in outcrossing plant species than self-pollinated plants, a higher mapping resolution is expected.

2.4.3 Germplasm

The germplasm plays a key role in LD variation because the extent of LD is influenced by the level of genetic diversity captured by the population under consideration. In general, the larger the genetic variation, the faster the LD decay, a direct consequence of the broader historical recombination. The population sample effect is evident in maize (*Zea mays*) where LD decays within 1 kb in landraces, approximately doubles (~ 2kb) in diverse inbred lines, and can extend up to several hundred kb in commercial elite inbred

lines (Jung et al. 2004). Tenaillon et al. (2001) investigated sequence diversity at 21 loci on chromosome 1 in a diverse group of maize germplasm, including exotic landraces and US accessions. An average LD decay was determined to occur within 400 bp ($r^2 = 0.2$), but extended up to 1000 bp in a group of US inbred lines. In Michigan local Arabidopsis populations, LD decay varied within 50 kb up to 50-100 cM. The latter was explained as a genetic bottleneck or founder effect, which reduced dramatically the genetic variation (Nordborg et al. 2002). In cotton (*Gossypium hirsutum*), the genome-wide average LD ($r^2 \leq 0.1$) declined to 10 cM in landraces, but was up to 30 cM in varieties (Abdurakhmonov et al. 2008). Myles et al. (2011) studied LD variation in over 1000 samples of domesticated grape (*Vitis vinifera*) and its wild relatives, reporting a rapid LD decay, even greater than in maize, as result of a weak domestication bottleneck followed by thousands of years of widespread vegetative propagation.

Estimates of genome-wide average LD decay may not reflect LD patterns between different populations of the same species. Each of these populations should be explored independently for the extent of LD in order to conduct successful association mapping studies (Abdurakhmonov and Abdukarimov 2008). Taking into account these three important biological factors, an obvious question is whether an increased or decreased level of LD is favourable in AM? Populations with either rapid or slow LD decay can be useful in AM, depending on the purposes of the study. Thus, populations with narrow genetic diversity and long extent of LD are amenable to coarse mapping with fewer markers requiring fine mapping in more genetically diverse populations, assuming that the causal genetic factors are sufficiently similar across different germplasm groups.

2.5 LD variation as effect of evolutionary factors

2.5.1 Selection

Simply stated, if alleles at two loci are in LD and they both affect positively reproductive fitness, the response to selection at one locus might be accelerated by selection affecting the other (Slatkin 2008). Thus, positive selection will increase LD between and in the vicinity of the selected loci, a phenomenon known as genetic hitchhiking. Even if the second locus is selectively neutral, the selection applied over the first will increase LD between them. The LD level between the two loci will remain constant over time depending on the genetic distance, the recombination rate and the effective population size (*N*). In contrast, if both loci are maintained by balancing selection, then LD can persist indefinitely (Lewontin 1964). Nonetheless, LD should be higher in loci affected by positive selection because a strong positive selection limits genetic diversity as opposed to a balancing selection which tends to maintain or increase polymorphism. In general, disease resistance genes in plants (*R*-genes) are affected by balancing selection with low intragenic LD and rapid decay (Yin et al. 2004), which could facilitate fine mapping of disease resistance genes. Domestication bottlenecks followed by strong selection for specific environments and end-use traits have modified the genome architecture in many crops reducing genetic diversity and creating population structure, which may be the main factor affecting the power of AM.

2.5.2 Population stratification

Selection affects the genome and LD in a locus-specific manner. In contrast, population stratification affects LD throughout the genome. Consequently, the power of AM can be

strongly reduced by population structure and family structure (Balding 2006). Population structure occurs from the unequal distribution of allele frequencies among subpopulations of different ancestries, while family structure occurs when individuals in a population display different degrees of relatedness (Würschum 2012). When these populations are sampled to construct an association mapping panel, the intentional or unintentional mixing of individuals with different allele frequencies and relatedness creates unlinked LD. Significant LD between unlinked loci results in false positive associations between a marker and a trait. Thornsberry et al. (2001) reported significant associations between polymorphisms at the maize *Dwarf8* gene and variation in flowering time, but they also stated that up to 80% of the false positive associations resulted from population structure. The occurrence of spurious associations is markedly higher in adaptation-related genes because they show positive correlations with the environmental variables under which they have evolved, and, as a result, the genomic regions carrying these genes could present stronger population differentiation. Several statistical models take into account the potential effect of population stratification. Commonly used algorithms are those of Pritchard and Rosenberg (1999) implemented in the software STRUCTURE (Pritchard et al. 2000; Hubisz et al. 2009), Principal Component Analysis (PCA) (Price et al. 2006) and kinship relationships (Yu et al. 2006).

2.5.3 Genetic drift, population bottleneck and gene flow

The effect of genetic drift in a small population results in the consistent loss of rare allelic combinations which increases LD level (Flint-Garcia et al. 2003). Genetic drift can create LD between closely linked loci. The effect is similar to taking a small sample from a

large population. Even if two loci are in linkage equilibrium, sampling only few individuals can create LD (Slatkin 2008).

LD can also be created in populations that have experienced a reduction in size (called a bottleneck) with accompanying extreme genetic drift (Dunning et al. 2000). After a bottleneck, some haplotypes will be lost; generally resulting in increased LD. Subsequent bottlenecks will further contribute to augment LD by increasing the effect of genetic drift. Colonizing species undergo repeated bottlenecks, and many models of the history of hominids assume the occurrence of a bottleneck when modern humans first left Africa (Noonan et al. 2006). Several studies of humans have argued that long distance LD in humans is the result of this early bottleneck in human history (Schmegner et al. 2005). In plants, comparisons with wild ancestors indicate that, in maize, approximately 80% of the allele richness has been lost as a consequence of domestication bottlenecks (Wright and Gaut 2005) while this number is 40-50% in sunflower (Liu and Burke 2006) and 10-20% in rice (Zhu et al. 2007). Gene flow introduces new individuals or gametes with different ancestries and allele frequencies among populations. If selection maintains differences in allele frequencies at two or more loci among subpopulations, LD in each subpopulation will persist (Slatkin 1975), but generally when random mating and recombination take place, LD caused by gene flow eventually breaks down.

Factors such as genetic drift, population bottlenecks and gene flow can contribute to generating artificial LD and negatively impact the ability to use LD in AM for the precise localization of QTL. In general, any biological or evolutionary forces that contribute to an increase of LD beyond that expected by chance in an “ideal” population will result in false-positive associations (Gaut and Long 2003).

2.6 Approaches for AM

Many methodologies have been developed and are widely used for AM in humans (Schulze and McMahon 2002), and several are perfectly applicable without change or with case-to-case modifications for a wide range of organisms, including plants. The methods to study marker-trait association using LD may differ for discrete and quantitative traits (Nielsen and Zaykin 2001). Approaches such as Multiparent Advanced Generation Intercross (MAGIC), Nested Association Mapping (NAM), Case-control (CC), Transmission Disequilibrium Test (TDT), genomic control (GC) and structured association (SA) are available.

2.6.1 Multiparent Advanced Generation Intercross (MAGIC)

MAGIC is an extension of the advanced intercross method in which an intermated mapping population is created from multiple founder lines. A Recombinant Inbred Line (RIL) population is created from multiple founder lines, in which the genome of the founders are first mixed by several rounds of mating, and subsequently inbred to generate a stable panel of inbred lines. The larger number of parental accessions increases the allelic and phenotypic diversity over traditional RILs, potentially increasing the number of QTL that segregate in the population. The control crosses allow breaking up the covariance between genotype and phenotype caused by population structure reducing the number of false positive QTL. In addition, MAGIC allows the identification of more stable QTL than those detected in biparental populations, which are often not transferable from one population to another (Holland 2007). The successive rounds of recombination cause LD to decay, thereby increasing the precision of QTL location (Mackay and Powell

2007). In both crops and animals, the MAGIC design has the ability to capture the majority of the variation available in the gene pool. Although it might take several years before these populations are suitable for fine mapping, they are relatively inexpensive to develop and their value as mapping resources increases with each generation (Mackay and Powell 2007). In plants, MAGIC can be used to combine coarse mapping with low marker densities on lines derived from an early generation, with fine mapping using lines derived from a more advanced generation and a higher marker density.

2.6.2 Nested Association Mapping (NAM)

Nested association mapping (NAM) design is another variant of connected crosses (Yu et al. 2008). In this design, a diverse set of parental lines is selected based on molecular diversity analysis, and each of the lines is crossed with a common reference line. Although NAM populations can reduce the confounding effects of population structure and increase the frequency of rare alleles, intercrossing the diverse panel of parental lines can better mitigate the drawbacks of classical AM populations (Guo et al. 2013).

2.6.3 Case-control (CC)

The classical methodology and design of AM is the "case and control" (CC) approach. If a mutation increases disease susceptibility, then we can expect it to be more frequent among affected individuals (cases) than among unaffected individuals (controls). The essential idea behind CC-based AM is that markers close to the disease mutation may also have allele frequency differences between cases and controls if there is LD between the marker locus and the "susceptibility" mutations (Schulze and McMahon 2002). For

accurate mapping, this design requires an equal number of unrelated and unstructured *case-control* samples. The Pearson χ^2 test, Fisher's exact test or Yates continuity correction can be used to compare allele frequencies and detect association between a phenotype and a marker (Abdurakhmonov and Abdukarimov 2008). The CC tests are sensitive to overall population LD between a marker and a locus affecting the trait. As previously discussed, LD can exist between unlinked loci, meaning that strong marker-trait association is not necessarily evidence for physical proximity between a marker and the gene affecting the phenotype. As a consequence, the CC approach is highly sensitive to population structure (Schulze and McMahon 2002). To efficiently eliminate the confounding effects caused by population structure, Spielman et al. (1993) developed the Transmission Disequilibrium Test (TDT).

2.6.4 Transmission Disequilibrium Test (TDT)

The ability to map QTL in collections of breeding lines, landraces or samples from natural populations has merit. In these populations, LD often decays more rapidly than in controlled crosses, enabling fine mapping. The challenge is to distinguish the effects of population subdivision from LD caused by linkage (syntenic LD). A robust method to test for this partitioning is the TDT (Spielman et al. 1993) that permits the detection of linkage in the presence of disequilibrium. Neither linkage alone nor disequilibrium alone (non syntenic LD) will generate a positive result in a TDT. As a consequence, the TDT is a robust method to control false positives (Mackay and Powell 2007). In brief, TDT compares the transmission versus the non-transmission of alleles to the offspring using a χ^2 test, assuming a linkage between a marker and a trait. The TDT design requires

genotyping of markers from three individuals: one heterozygous parent, one homozygous parent and one affected offspring. In the absence of linkage between QTL and marker, the expected ratio of transmission to non-transmission is 1:1 (Nielsen and Zaykin 2001). In the presence of linkage, it is distorted to an extent that depends on the strength of LD between the marker and the QTL. In addition, the power of the association will depend on the effectiveness of selection of extreme progeny in driving segregation away from expectation (Mackay and Powell 2007).

The initial TDT approach did not address the cases of multiallelic markers, multiple markers, missing parental information, large pedigrees and complex quantitative traits (Schulze and McMahon 2002). A variety of extensions of the TDT approach have been developed and applied to resolve multiallelic marker issues (i.e., GTDT, ETDT, MCTm); reviewed by Schulze and McMahon (2002).

In crops, parental and progeny lines are often separated by several generations of gametogenesis rather than one, as is often the case of human studies. For this reason, the TDT, while still valid, may be less robust because the breeding process may result in increased segregation distortion (Mackay and Powell 2007).

2.6.5 Other approaches

Population structure arising from recent migration, population admixture and artificial selection will generate unlinked LD. Assuming that such population structure has a similar effect on all loci, a random set of markers can be used to statistically assess the extent with which population structure is responsible for unlinked LD (Stich and Melchinger 2010). This is the basis of genomic control (GC). For example, for a case-

control analysis of candidate genes, the GC approach computes χ^2 test statistics for independence for both null (random) and candidate loci. An average χ^2 of null loci greater than 1.0 indicates the presence of significant structure. By using the magnitude of the χ^2 test observed at the null loci, a multiplier is derived to adjust the critical value for significance tests for candidate loci (Mackay and Powell 2007). By contrast, structure association (SA) analysis developed by Pritchard et al. (2000), first uses a set of random markers to estimate population structure (*Q* matrix), and then incorporates this estimate into a general linear model (GLM) analysis which enables correction for false associations. Yu et al. (2006) developed a mixed linear model (MLM), which incorporates both population structure and familial relatedness (*K* matrix). Their simultaneous use, however, may result in an over-correction and consequently in a reduced QTL detection power if *Q* and *K* explain a major part of the phenotypic variation (Würschum 2012). Another type of mixed model incorporates principal component analysis (PCA) instead of the *Q* matrix (Price et al. 2006). The PCA-based MLM model is computationally effective as compared to the *Q* matrix estimated from STRUCTURE. Studies conducted in humans, Arabidopsis and bread wheat (Raman et al. 2010; Yu et al. 2006; Zhao et al. 2007) have demonstrated the effectiveness of the MLM approach over the GLM. The application of the *K* matrix has recently been shown to be sufficient for the analysis of breeding populations (Bradbury et al. 2011; Wang et al. 2011).

2.7 AM studies in plants

Some of the first LD mapping studies in plants were done in maize (*Zea mays*) (Bar-Hen et al. 1995), rice (*Oryza sativa*) (Virk et al. 1996) and oat (*Avena sativa*) (Beer et al.

(1997). Bar-Hen et al. (1995) and Virk et al. (1996) predicted the association of quantitative traits using RAPD and isozymes markers, respectively. Beer et al. (1997) associated 13 QTL with RFLP loci using 64 oat varieties and landraces. In these studies, a low number of genome-wide distributed markers were assessed without considering population stratification. The first empirical candidate gene association taking into account background molecular markers to correct for population structure was performed in maize looking at the D8 locus and its association with flowering time (Pritchard 2001). In Arabidopsis, most of the AM studies were focused on providing proof of concept, identification of QTL involved in adaptation and detection of additional alleles to supplement other mutagenesis approaches (Ersoz et al. 2007). Aranzana et al. (2005) performed the first attempt at a genome wide association study (GWAS) in Arabidopsis, reporting previously known flowering time and three known pathogen-resistance genes. GWAS refers to the use of many markers that span an entire genome to identify functional common variants in LD with at least one of the genotyped markers. Numerous research papers focusing on LD and AM have since been published on more than a dozen plant species. These studies have been reviewed by Gupta et al. (2005) and more recently by Zhu et al. (2008).

In the last five years, plant AM studies have expanded because of advances in sequencing technologies which enable more efficient and cost-effective development of a large number of molecular markers such as Single Nucleotide Polymorphisms (SNPs). In Arabidopsis, new studies have been carried out aiming to dissect downy mildew resistance genes and climate-sensitive QTL, with special efforts focused on the understanding of adaptive variation (Li et al. 2010; Nemri et al. 2010). The first applied a

CG approach, and the second a GWAS based on no fewer than 213,497 SNPs. In maize, recent studies dissected the quantitative genetic nature of the northern leaf blight (NLB) resistance, southern leaf blight (SLB) resistance and leaf architecture, scanning the genome using ~1.6 million SNPs (Kump et al. 2011; Poland et al. 2011; Tian et al. 2011). Poland et al. (2011) identified several loci with small additive effects carrying candidate genes related to plant defense, including receptor-like kinase genes. Kump et al. (2011), from the same research group, identified 32 QTL with predominantly small additive effects related to SLB resistance. Similarly, Tian et al. (2011) demonstrated that the genetic architecture of leaf traits is dominated by small effects and that the *liguleless* genes have contributed to more upright leaves. Li et al. (2013) using ~1 million SNPs dissected the genetic architecture of oil biosynthesis in maize kernel identifying 74 loci associated with kernel oil and fatty acid composition. Currently, whole genome scanning has moved beyond Arabidopsis and maize to other species such as rice and barley. Huang et al. (2010a) uncovered the genetic basis of 14 rice agronomic traits based on ~3.6 million SNPs. The loci identified through GWAS explained ~36% of the phenotypic variance on average, and 32 new loci associated with flowering time and grain-related traits were identified. In barley, GWAS of 15 morphological traits identified one putative anthocyanin pathway gene, *HvbHLH1*, carrying a deletion resulting in a premature stop codon and which was diagnostic for the absence of anthocyanin in the germplasm studied (Cockram et al. 2010). Efforts towards understanding adaptation-related genes have been undertaken in wheat. Raman et al. (2010) applied GWAS in order to identify genetic factors associated with aluminium resistance, one of the most restrictive abiotic stresses on acid soils worldwide. The study confirmed previously identified loci and identified

putative novel ones. Subsequently, Rousset et al. (2011) studied the genetic nature of flowering time in wheat to investigate the effect of candidate genes on flowering time. The *Vrn-3* gene explained a high percentage of the phenotypic variation of earliness followed to a lesser extent by *Vrn-1*, *Hd-1* and *Gigantea* (*GI*). In *Brassica napus*, several seed oil related loci were identified, with a few corresponding to previously reported genomic regions associated with oil variation (Zou et al. 2010). In tetraploid alfalfa (*Medicago sativa*), fifteen SSR markers showed strong association with yield in different environments (Li et al. 2011a). In sugar beet (*Beta vulgaris*), genetic variation of six agronomic traits was dissected using GWAS, identifying several QTL with major effects and others with epistatic effects (Würschum et al. 2011). Thus, LD mapping, considered a few years ago as an emerging tool in plant genomics, has recently been shown to be a powerful method to dissect complex traits in crops. Appendix I summarizes these and other recently published AM studies in plants. Earlier publications are summarized elsewhere (Gupta et al. 2005; Zhu et al. 2008).

2.8 Benefits and limitations of AM

The potential high resolution in localizing a QTL controlling a trait of interest is the primary advantage of AM as compared to linkage mapping (Fig. 2.3). AM has the potential to identify more and superior alleles and to provide detailed marker data in a large number of lines which could be of immediate application in breeding (Yu and Buckler 2006). Furthermore, AM uses breeding populations including diverse and important materials in which the most relevant genes should be segregating. Complex interactions (epistasis) between alleles at several loci and genes of small effects can be

identified, pinpointing the superior individuals in a breeding population (Tian et al. 2011). Sample size and structure do not need to be as large as for linkage studies to obtain similar power of detection. Finally, AM has the potential not only to identify and map QTL but also to identify causal polymorphisms within a gene that are responsible for the difference between two phenotypes (Palaisa et al. 2003). AM suffers from some limitations such as when the trait under consideration is strongly associated with population structure. Most traits under local adaptation or in balancing selection in different populations may be thus affected (Stich and Melchinger 2010). When statistical methods to correct for population structure are applied, the differences between subpopulations are disregarded when searching for marker-trait associations. Therefore, all polymorphisms responsible for the phenotypic differences between subpopulations remain undetected, thus underpowering AM. The drawback of population structure, however, could be mitigated if AM analyses are conducted separately within subpopulations (Zhao et al. 2011). LD mapping often requires a large number of markers for genotyping in GWAS. The number of markers depends in large part on the genome size and the expected LD decay; linkage mapping generally requires fewer markers to detect significant QTL. A high density of markers can only be achieved through the development of an integrated genotyping by sequencing platform. Thus, the analysis of cost-benefit must be conducted in the light of the real impacts that such investments will have in the future market appreciation of that plant species. Alternative approaches such as linkage mapping and CG could be feasible for other studied traits. The power of AM to detect an association is influenced by allele frequency distribution at the functional polymorphism level.

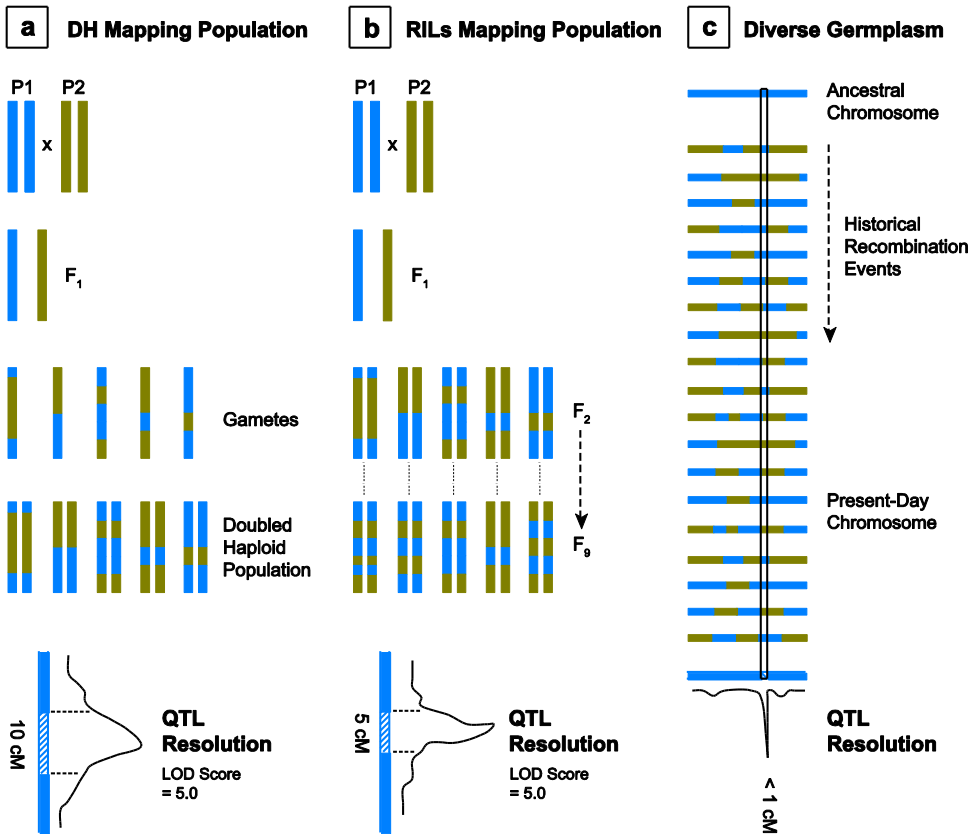


Fig. 2.3 Comparison of mapping resolution between linkage mapping and AM. **a** A Doubled Haploid (DH) mapping population. **b** A Recombinant Inbred Line (RIL) mapping population. **c** A collection of diverse germplasm. **a** and **b** present low QTL resolution as a consequence of few meiosis events accumulated, **c** presents a high QTL resolution because a larger number of recombination events have accumulated during the population history.

The results of empirical studies suggest that a high percentage of alleles are rare (Myles et al. 2009). Rare alleles cannot be evaluated adequately because, by definition, they are present in too few individuals and consequently lack resolution power. As a consequence, an important piece of heritability remains undetected. For such rare alleles, linkage mapping and family based AM may be used because correlation between population structure and phenotypes can be broken, and allele frequencies can be inflated to enhance the power of mapping (Stich and Melchinger 2010). In this regard, several studies have combined linkage mapping and AM mapping, which reduces spurious associations caused by population structure, particularly for traits strongly affected by

local geographic patterns (Brachi et al. 2010; Poland et al. 2011; Cadic et al. 2013). With the growing interest in finding the missing heritability not accounted for by common alleles (Asimit and Zeggini 2010), several new association analysis methods for rare variants are being proposed, with some important advances in complex trait dissection (Li and Leal 2008). The simplest approach is to test them individually using standard contingency table and regression methods such as those implemented in the genetic software PLINK (Purcell et al. 2007). This method, called "single-locus test" is highly problematic, given, for example, the poor power that such statistical tests have to detect small differences between diagnostic or phenotypic groups (Gorlov et al. 2008). Other methods that overcome the power issues associated with testing rare variants individually include the collapsing strategy, methods based on summary statistics, multiple regression and data mining which are comprehensively reviewed by Bansal et al. (2010). Approaches involving direct sequencing have been tested by Li and Leal (2008). Since epigenetic factors are also likely to contribute to common complex traits, epigenome-wide association studies (EWASs) have been proposed to uncover another missing piece of heritability as yet unexplained by common variants (Rakyan et al. 2011), specifically involving the study of variation in DNA methylation across the genome.

2.9 Computer programs for AM

A variety of software packages are available for AM. TASSEL is a commonly used software for LD mapping in plants, frequently updated with newly developed methods. TASSEL can be used for AM analysis using GLM and MLM, calculation and graphical display of LD statistics, analysis of population structure using PCA and tree plots of

genetic distance. Although TASSEL can handle both SSR and SNP markers, the latest version only accepts SNPs. For SSR analysis, users must continue with TASSEL v. 2.1. Alternatively, GenStat v. 15 offers traditional statistical analyses as well as linkage and AM analyses for SSRs. GenStat v. 15 performs structure analysis based on PCA, LD decay and single trait association analysis using PCA-based MLM. Version 15 was recently released (http://www.vsni.co.uk/webstore/software/genstat). Gupta et al. (2005) and Excoffier and Heckel (2006) comprehensively reviewed the most common software for population genetics and LD mapping analyses but the majority of them can only handle a few thousand marker loci. Progress in sequencing technologies has solved the past issue of genotyping large populations with high marker densities and software development has also moved quickly. Nowadays, the main issue is the time required for processing large data sets and the availability of powerful statistical models to adjust for multiple testing.

JMP Genomics v.6 is a Windows based program that offers several solutions for handling large SNP data sets (http://www.jmp.com/software/genomics). Among its main characteristics, JMP Genomics is capable of handling data sets as large as 1.5 million SNPs for 15,000 samples on a 32-bit desktop work station using CG or GWA. It also corrects for relatedness and population structure using association tests, and calculates identical by descent (IBD), identical by state (IBS) and allele-sharing individual relationship matrices. Interactive triangular plots and zooming features permit visualization of LD blocks. Association between SNPs and multiple traits can be tested separately or jointly, while adjusting for covariates. JMP Genomics v. 6 also simplifies the analysis of rare and common variants, and includes features for high quality graphs

and figures. Similar applications can be found in the GenAMap software, which incorporates visualization strategies for structured AM (http://cogito-b.ml.cmu.edu/genamap/). It has a processing capacity of 1 million SNPs in approximately 1 hour. The analysis is performed on a remote cluster complete with complex parallelization schemes to optimize run-time efficiency. GenAMap gives an overview of the association results through a heatmap view where SNPs are plotted against a network of candidate genes, shows interactions between genes, integrates the association strengths of the genes to SNPs in the genome, and creates a tree view of structured genes to explore and identify functional relevant branches of the tree that are associated with a genomic region. Although GenAMap was primarily developed for human diseases, it can be applied to plant AM as well. PLINK software v. 1.07 (Purcell et al. 2007; http://pngu.mgh.harvard.edu/purcell/plink/) is an open source C/C++ GWAS tool set. With PLINK, large data sets comprising hundreds of thousands of SNPs and individuals can be readily manipulated and analyzed. PLINK offers five main characteristics. Data management is a simple interface for reordering, recording and filtering genotypic information. Summary statistics to determine the randomness of genotyping failure highlights the test of missingness on a simple haplotypic case-control test. Population stratification is measured on the basis of a genome average proportion of alleles sharing identical by state (IBS) between any two individuals. PLINK offers tools to cluster individuals into homogeneous subsets to identify potential outlier individuals causing genotyping or pedigree errors, and to incorporate this stratification in GWAS. Association analyses include CC, stratified analysis, TDT, QTDT, sib TDT and

correction for multiple tests. Appendix II summarizes these and other software based on their analytical focus.

GENETIC CHARACTERIZATION OF A CORE COLLECTION OF FLAX (*LINUM USITATISSIMUM* L.) SUITABLE FOR ASSOCIATION MAPPING STUDIES AND EVIDENCE OF DIVERGENT SELECTION BETWEEN FIBER AND LINSEED TYPES

Braulio J. Soto-Cerda[1,2] · Axel Diederichsen[3] · Raja Ragupathy[1,2] · Sylvie Cloutier[1,2]

[1]University of Manitoba, Department of Plant Science, 66 Dafoe Road, Winnipeg, MB, R3T 2N2, Canada

[2]Cereal Research Centre, Agriculture and Agri-Food Canada, 195 Dafoe Rd, Winnipeg, MB, R3T 2M9, Canada

[3]Plant Gene Resources of Canada, Agriculture and Agri-Food Canada, 107 Science Place, Saskatchewan, SK, S7N 0X2, Canada

Author Braulio J. Soto-Cerda carried out the analyses, interpretation of data and co-wrote the manuscript. The major supervisor Dr. Sylvie Cloutier, designed the study, generated the data, supervised the work and co-wrote the manuscript. Dr. Axel Diederichsen characterized and developed the flax core collection. Dr. Raja Ragupathy carried out the gene prediction and gene annotation.

The manuscript was published in **BMC Plant Biology 2013, 13:78.**

3.0 GENETIC CHARACTERIZATION OF A CORE COLLECTION OF FLAX (*LINUM USITATISSIMUM* L.) SUITABLE FOR ASSOCIATION MAPPING STUDIES AND EVIDENCE OF DIVERGENT SELECTION BETWEEN FIBER AND LINSEED TYPES

3.1 Abstract

Flax is valued for its fiber, seed oil and nutraceuticals. Recently, the fiber industry has invested in the development of products made from linseed stems, making it a dual purpose crop. Simultaneous targeting of genomic regions controlling stem fiber and seed quality traits could enable the development of dual purpose cultivars. However, the genetic diversity, population structure and linkage disequilibrium (LD) patterns necessary for association mapping (AM) have not yet been assessed in flax because genomic resources have only recently been developed. We characterized 407 globally distributed flax accessions using 448 microsatellite markers. The data was analyzed to assess the suitability of this core collection for AM. Genomic scans to identify candidate genes selected during the divergent breeding process of fiber flax and linseed were conducted using the whole genome shotgun sequence of flax. Combined genetic structure analysis assigned all accessions to two major groups with six sub-groups. Population differentiation was weak between the major groups (F_{ST} = 0.094) and for most of the pairwise comparisons among sub-groups. The molecular coancestry analysis indicated weak relatedness (mean = 0.287) for most individual pairs. Abundant genetic diversity was observed in the total panel (5.32 alleles per locus), and some sub-groups showed a high proportion of private alleles. The average genome-wide LD (r^2) was 0.036, with a

relatively fast decay of 1.5 cM. Genomic scans between fiber flax and linseed identified candidate genes involved in cell-wall biogenesis/modification, xylem identity and fatty acid biosynthesis congruent with genes previously identified in flax and other plant species. Based on the abundant genetic diversity, weak population structure and relatedness and relatively fast LD decay, we concluded that this core collection is suitable for AM studies for targeting multiple agronomic and quality traits aiming at the improvement of flax as a true dual purpose crop. Our genomic scans provide the first insights of candidate regions affected by divergent selection in flax. In combination with AM, genomic scans have the ability to increase the power to detect loci influencing complex traits.

3.2 Introduction

Flax (*Linum usitatissimum* L.) is an annual, self-pollinated species with a genome size of ~ 370 Mb (Ragupathy et al. 2011). The species is believed to have originated in either the Middle East or Indian regions (Vavilov 1951) and spread throughout Asia and Europe, prior to its introduction into the New World (Green et al. 2008). Divergent selection applied over thousands of years has resulted in fiber and linseed types which are the same species but differ considerably in morphology, anatomy, physiology and agronomic performance (Diederichsen and Ulrich 2009). Fiber flax cultivars are taller and less branched and are grown in the cool-temperate regions of China, the Russian Federation and Western Europe (Green et al. 2008). Linseed cultivars are shorter, more branched, larger seeded and are grown over a wider area in continental climate regions such as Canada, India, China, the United States and Argentina (Green et al. 2008). Flax provides

raw materials for food, medicine and textiles and, as such, it has been of great importance to human culture and development for more than 8,000 years (van Zeist and Bakker-Heeres 1975). Linseed oil is well-known for its health benefits mainly attributed to its high content of omega-3 alpha linolenic acid (55-57%). Linseed oil has been used for centuries in paints and varnishes because of its unique drying properties attributable to its distinctive fatty acid composition (Przybylski 2001). Consumption of ground seeds adds nutritional benefits because flax seeds are also a rich source of lignans, compounds that have anticancer properties (Westcott and Muir 2003). In the last decade, the fiber industry has devoted some effort to develop high-value products from linseed stems with applications in the pulp, technical fiber and biofuel industries (Diederichsen and Ulrich 2009; Cullis 2011). Therefore, understanding the genetic diversity of flax collections is important for the continued improvement of this crop as well as for its development into a truly dual purpose crop (Cullis 2011).

Initial diversity assessments in flax were carried out using morphological parameters (von Kulpa and Danert 1962; Diederichsen 2001; Diederichsen et al. 2006; Diederichsen and Raney 2006) and isozymes (Tyson et al. 1985; Månsby et al. 2000). In recent years, molecular marker systems such as randomly amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), inter-simple sequence repeat (ISSR), simple sequence repeat (SSR) and inter-retrotransposon amplified polymorphism (IRAP) have been used to measure genetic variation and relationships in cultivars and landraces of flax (Spielmeyer et al. 1998; Everaert et al. 2001; Fu et al. 2002, 2003; Wiesnerová and Wiesner 2004; Fu 2005, 2011; Diederichsen and Fu 2006, 2008; Cloutier et al. 2009; Rajwade et al. 2010; Uysal et al. 2010; Smỳkal et al. 2011; Soto-Cerda et al.

2012; Cloutier et al. 2012a). However, most of these previous studies assessed either few marker loci or few genotypes.

World gene banks store approximately 48,000 accessions of flax germplasm (Diederichsen 2007). In Canada, a world collection of approximately 3,500 accessions of cultivated flax is maintained by Plant Gene Resources of Canada (PGRC). This collection has traditionally been deployed in flax breeding through a variety of conventional strategies (Green et al. 2008). In 2009, the Total Utilization Flax GENomics (TUFGEN; http://www.tufgen.ca) project was initiated in Canada to generate genomics resources for flax and to apply them to an array of traits for the ultimate purpose of flax improvement. The TUFGEN project has developed numerous genomics resources including molecular markers (Cloutier et al. 2009, 2012a; Kumar et al. 2012), genetic maps (Cloutier et al. 2011, 2012b), a physical map and bacterial artificial chromosome end sequences (Ragupathy et al. 2011), expressed sequence tags (Venglat et al. 2011) and whole genome shotgun sequence (Wang et al. 2012a). To take advantage of these tools, a core collection of 407 flax accessions capturing the breadth of the phenotypic diversity of the PGRC collection was assembled.

Quantitative trait loci (QTL) and association mapping (AM) are complementary approaches for the identification of marker-trait association. The first utilizes biparental mapping populations to monitor the co-segregation of QTL and marker loci. The second utilizes germplasm collections to identify QTL-marker correlations based on LD (Flint-Garcia et al. 2003). QTL analysis has limited mapping resolution due to the accumulation of few meiosis events in a single cross, but it is not affected by population structure which can be a source of spurious association in AM. Conversely, AM can achieve

higher mapping resolution through high numbers of historical recombination events in germplasm collections. An ideal association panel should harbor the broadest genetic diversity because this is often correlated with a rapid LD decay necessary to resolve complex trait variation(s) to a single gene or nucleotide (Myles et al. 2009). Null or weak population structure and a low level of relatedness among individuals of the germplasm collection are also desirable. Thus, genetic diversity, population structure, familial relatedness and LD patterns need to be assessed prior to AM analysis to fully exploit their advantages for flax genetic improvement.

In this study, we genotyped 407 flax accessions using 448 microsatellite loci. The overall goal was to evaluate the usefulness of this flax core collection for AM studies. Our specific goals were (1) to investigate the genetic diversity; (2) to estimate the levels of population structure and assess familial relatedness; (3) to detect the patterns of LD; and (4) to identify non-neutral genomic regions potentially underlying divergent selection between fiber and linseed types.

3.3 Material and methods

3.3.1 Plant material

The PGRC flax collection has been evaluated in the field to measure seed characteristics, disease resistance and phenological traits (Diederichsen and Fu 2008). Based on this information, a core collection of 381 flax accessions was assembled representing the phenotypic diversity of the PGRC flax world collection. To these, 26 accessions of relevance to recent Canadian flax breeding programs were added, resulting in a core collection of 407 accessions. Information on the geographic origin and improvement

status of the accessions is shown (Appendix III, IV). The core collection comprised 92 fiber accessions, 285 linseed accessions and 30 unknown accessions.

3.3.2 DNA isolation and microsatellite genotyping

Genomic DNA was extracted from leaf tissues collected from a single plant of each accession (Cloutier et al. 2009). DNA was quantified using a fluorometer and diluted to a 6 ng/µL working solution. Four hundred forty eight microsatellites (Cloutier et al. 2009, 2012a; Roose-Amsaleg et al. 2006; Deng et al. 2010, 2011) distributed across the 15 linkage groups (Cloutier et al. 2012b) were analyzed following the procedure previously described (Cloutier et al. 2009). Briefly, the amplification products were resolved on an ABI 3130xl Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). Output files were analyzed by GeneScan (Applied Biosystems) and subsequently imported into Genographer. Fragment sizes were estimated using the GeneScan ROX-500 and MapMarker® 1000 (BioVentures Inc., Murfreesboro, TN) internal size standards, and the genotypic data matrix generated was used for all posterior analyses. The genotype of each locus was encoded based on its allele size in bp or as a null allele for dominant markers. The selective neutrality status was tested across microsatellites prior to other downstream genetic analyses using the Ewens-Watterson (EW) neutrality test (Manly 1985) implemented in POPGENE v.1.31 (Yeh et al. 1997) with 1,000 permutations without replacement.

3.3.3 Phylogenetic analysis

To assess the genetic relationships among the accessions of the core collection, a dendrogram was generated using the neighbour-joining (NJ) algorithm based on the Nei (1973) minimum genetic distance method implemented in PowerMarker v.3.25 (Liu and Muse 2005) and displayed by MEGA 5 (Tamura et al. 2011). Nei (1973) minimum genetic distance method is applicable to any population without regard to the number of alleles per locus, the pattern of evolutionary forces and the reproductive method of the organism studied. Thus it is a realistic estimation of the genetic relationships in an artificial population when individuals display different selection intensities, breeding objectives, as well as improvement status. The analysis was performed with the 414 neutral microsatellites identified by the EW neutrality test including minor allele frequency (MAF) < 0.05. The genotype of each marker was encoded as two alleles using their sizes estimated above as follow: homozygous state (allele1/allele1) and heterozygous state (allele1/allele2). Null alleles "null/null" were encoded as 999/999 and missing values as "?/?". The reliability of the dendrogram topology was confirmed with 1,000 bootstraps with replacements.

3.3.4 Population structure

To investigate the patterns of population structure, we conducted principal coordinate (PCoA) and Bayesian-based analyses. Because LD can affect both PCoA and STRUCTURE analyses, we thinned the marker set by excluding microsatellites in strong LD, i.e., markers with a square of the correlation coefficient (r^2) greater than 0.4 (Yunusbayev et al. 2011). The frequency of the alleles was calculated in PowerMarker

v.3.25 (Liu and Muse 2005) and MAF < 0.05 were set to "U" (missing data) and excluded from the LD analysis. Genetic distances between markers were obtained from the microsatellite consensus linkage map of flax (Cloutier et al. 2012b) integrated with the physical map (Ragupathy et al. 2011). Linked and unlinked LD (r^2) was determined using GGT 2.0 (van Berloo 2008) with genotypic data encoded as follows: 100/100 = A, 200/200 = B, 300/300 = C and so on, where each letter represents a different allele. Heterozygous individuals were considered missing value "U".

PCoA was performed in a multidimensional space with data standardization using GENALEX v.6.41 (Peakall and Smouse 2006). Population structure analysis was carried out using STRUCTURE 2.3.3 (Pritchard et al. 2000; Hubisz et al. 2009). The admixture model was used with a burn in of 10,000 and 100,000 iterations for *K* populations ranging from 1 to 12. Ten runs for each *K* value were performed and the *ad-hoc* statistic Δ*k* was used to determine the optimum number of sub-groups (Evanno et al. 2005). Prior to population structure analysis, SSR data was encoded using the size of each allele and "-9" was used for missing values. Accessions with estimated memberships ≥ 0.70 were assigned to corresponding groups; accessions with estimated memberships < 0.70 were assigned to a mixed group. We adopted a cut-off value of 0.70 because 85% of the accessions are cultivars and breeding material, thus it is likely that their genome structure resembles more than one ancestral population. The inferred sub-groups were visualized in Distruct (Rosenberg 2004). Pairwise F_{ST} comparisons were calculated using GENALEX v.6.41 (Peakall and Smouse 2006) to determine the genetic differentiation between the inferred genetic groups.

3.3.5 Molecular coancestry

Strong familial relatedness can potentially inflate the number of spurious associations when it is not accounted for by the AM model. Relatedness was estimated using the molecular coancestry parameter (f_{ij}) according to Caballero and Toro (Caballero and Toro 2002). The molecular coancestry between two individuals *i* and *j* is the probability that two randomly sampled alleles from the same locus in two individuals are identical by state (Caballero and Toro 2002). Molecular coancestry between two individuals *i* and *j* at a given locus can be computed using the following scoring rules (Caballero and Toro 2002): $fij_{,l} = ¼[I_{11} + I_{12} + I_{21} + I_{22}]$, where I_{xy} is 1 when allele *x* on locus *l* in individual *i* and allele *y* in the same locus in individual *j* are identical and zero otherwise. Notice that this estimate can only have four values: 0, ¼, ½, and 1. The molecular coancestry between two individuals *i* and *j* (*fij*) can be obtained simply by averaging over *L* analyzed loci. Molecular coancestry matrices comparing all pairs of individuals within the core collection and within the different genetic groups identified above were calculated using all 448 microsatellites using MolKin v.3.0 (Gutierréz et al. 2005). Genotypic data based on the size of alleles was encoded as two alleles following the Genpop software format as follows: 100/200 = 0102, 200/200 = 0202 and so on. Missing values were labeled "0000".

3.3.6 Genetic diversity

Genetic diversity parameters were estimated across the genetic groups identified above based on the 414 neutral microsatellites. Unbiased gene diversity (UH_e), observed heterozygosity (H_o), total number of alleles (N_a), inbreeding coefficient (F_{IS}) and

polymorphic loci (%) were calculated in GENALEX v.6.41 (Peakall and Smouse 2006). Allelic richness (R_s) and private alleles ($\prod$) were corrected for sample size differences and estimated using the rarefaction method implemented in HP-RARE v.1.2 (Kalinowski 2005). The number of rare alleles (MAF < 0.05) and the polymorphism information content (PIC) values were calculated in PowerMarker v.3.25 (Liu and Muse 2005).

3.3.7 Linkage disequilibrium

LD was estimated by calculating r^2 using GGT 2.0 (van Berloo 2008) as described in the population structure section above. Only microsatellites with known chromosome information in the consensus map of flax (Cloutier et al. 2012b) were used for LD estimation. Microsatellites on the same linkage group were considered linked and those on different linkage groups, unlinked. Mean LD was estimated for linked and unlinked markers in the total panel and for the different genetic groups identified by NJ and population structure analyses. The 95^{th} percentile of r^2 distribution for unlinked markers was considered the cut-off LD value to determine whether LD resulted from physical linkage (Breseghello and Sorrells 2006). Average genome-wide LD decay versus genetic distance was estimated as previously described (Breseghello and Sorrells 2006). A cut-off value of $r^2 = 0.1$ was set to estimate the average genome-wide LD block. In order to compare the trend of LD decay amongst the different genetic groups, we averaged LD values to distance intervals equal to the average genome-wide LD block estimated.

3.3.8 Identification of non-neutral loci

To identify candidate loci linked to genomic regions that might have experienced divergent selection, we used the 92 fiber flax accessions present in the core collection (Appendix III). The "line selection" module in PowerMarker v.3.25 (Liu and Muse 2005) allows the selection of a core set of lines from a large germplasm collection that maximizes the genetic diversity. Likewise, this module enables the selection of a random set of lines from a large population. Using PowerMarker v.3.25 (Liu and Muse 2005) we randomly selected a set of 92 linseed accessions (among the 285 linseed accessions of the core collection) that captured the average number of alleles present in 100 random sets of 92 lines for the identification of non-neutral loci. Because bottlenecks can create false positive outliers, both fiber and linseed groups were analyzed with BOTTLENECK v.1.2.02 assuming the two-phase mutation model proposed for microsatellite data (Cornuet and Luikart 1996). Genotypic data format followed the Genepop format described above. We applied four outlier tests to minimize the number of false positives. (1) The Ewens-Watterson (EW) test statistic which identifies positively selected loci by evaluating significant deviation from expected heterozygosity (Dh/sd) in a single population (Watterson 1978) was calculated using BOTTLENECK v.1.2.02 (Cornuet and Luikart 1996). Statistical significance ($Dh/sd < -2.5$, $P < 0.05$) was assigned based on 1,000 permutations without replacement. (2) The ln RH test that identifies loci that differ in variability from the remainder of the genome by calculating the ratio of gene diversity in two populations was performed (Kauer et al. 2003). After standardization of ln RH estimates, 95% of the neutral loci are expected to have values ranging between -1.96 and 1.96. Any locus with a value higher than 1.96 ($P < 0.05$) was considered non-neutral. (3)

The Beaumont and Nichols (1996) approach implemented in LOSITAN (Antao et al. 2008) identifies loci under selection based on the distribution of heterozygosity and F_{ST} under an island model of migration. The expected null distribution of F_{ST} values and estimated P values for each locus were obtained (Antao et al. 2008). Loci exceeding the 95% upper confidence area were considered non-neutral. Genotypic data also followed the Genepop format described above. (4) The hierarchical island model that identifies outlier loci by allowing the exchange of more migrants within groups than between groups while generating the null distribution of F_{ST} values to reduce the number of false positives, was also applied to the data set (Excoffier et al. 2009). The fiber and linseed groups were analyzed with STRUCTURE 2.3.3 (Pritchard et al. 2000; Hubisz et al. 2009) to determine the number of groups to incorporate in the hierarchical analysis using the *ad-hoc* statistic Δk (Evanno et al. 2005). The expected F_{ST} distributions were obtained using Arlequin v.3.5 (Excoffier and Lischer 2010). Loci outside the 95% upper confidence area were considered non-neutral ($P < 0.05$). The genotype of each marker was encoded as two alleles using their size estimate where the homozygous state was 100100 and the heterozygous state was 100200. Null alleles "null/null" were encoded as 999999 and missing values were "??". Loci identified by at least two of the above four tests were retained and investigated as candidates for divergent selection.

3.3.9 Candidate genes

To identify candidate genes by homology search, we used the combined information of the consensus genetic map (Cloutier et al. 2012b), the physical map (Ragupathy et al. 2011) and the whole genome shotgun (WGS) sequence assembly (Wang et al. 2012a;

http://www.phytozome.net) of flax. When the candidate locus and its adjacent marker with the highest LD ($r^2 > 0.4$) were located in the same WGS sequence assembly scaffold, we estimated the physical to genetic distance (Mb/cM) to define the physical distance to be investigated for the identification of candidate genes. When adjacent markers were on different scaffolds or showed weak LD ($r^2 < 0.2$), we limited the search for candidate genes to the 10 kb regions upstream and downstream of the outlier markers. Annotation of the WGS assembly using the Hidden Markov Model-based gene-finding programs Augustus v.2.5.5 (Stanke et al. 2008) and GlimmerHMM v.3.0.1 (Majoros et al. 2004) were used. Using the BLASTn algorithm, predicted open reading frames of candidate genes were searched against an in-house flax EST database comprising 462,190 flax ESTs (Cloutier et al. 2009; Fenart et al. 2010; Venglat et al. 2011; NCBI *Linum usitatissimum* ESTdb) for evidence of expression, using an E-value cutoff of 1e-5. The same candidate gene sequences were used to perform BLASTx searches against the 16 million annotated proteins in the UniProtKB db (The UniProt Consortium 2009) to provide evidence of protein function using an E-value cutoff of 1e-5. Gene ontology (GO) annotations (Ashburner et al. 2000; http://www.geneontology.org) were also retrieved from the UniProtKB. Plant GO-slims for all three independent GO categories namely, cellular components, molecular functions and biological processes were obtained from all GO terms associated with the BLASTx gene annotations using the GO-slim viewer from the AgBase web server (McCarthy et al. 2006; http://www.agbase.msstate.edu).

3.4 Results

3.4.1 Phylogenetic analysis

Based on 414 neutral loci, the phylogenetic analysis using the NJ algorithm partitioned the 407 accessions into two major groups and one admixed group (Figure 3.1a; Appendix V). Group 1 (G1) was composed of 155 accessions that were further subdivided into three sub-groups representing accessions from South Asia, Western Europe and South America. The South Asian sub-group included mostly accessions from India, Pakistan and Afghanistan while the Western European sub-group contained mostly accessions from France, Portugal and Germany but also from Romania and Turkey. The South American sub-group included accessions from Argentina and Uruguay. Group 3 (G3) had 206 accessions distributed into two sub-groups, namely North America and Eastern Europe. The North American sub-group clustered cultivars and breeding materials originating exclusively from Canada and the U.S.A. However, not all North American accessions clustered within that sub-group. A few of these accessions were included in the Eastern European sub-group which otherwise included mostly accessions from Russia, Ukraine, Romania, Poland and Lithuania. This sub-group included 90% of the fiber flax accessions present in the core collection. Within the Eastern European sub-group, the geographic origin and industrial use overlapped, including fiber flax accessions from the Netherlands, the former Soviet Union and the U.S.A. The admixed group (G2) namely the North American/European group had 43 accessions from the U.S.A., Canada and European countries.

3.4.2 Population structure

A total of 259 loci meeting the neutrality criteria, LD < 0.4 and distributed across the 15 linkage groups, were included in these analyses. Similar to the phylogenetic analysis based on the NJ algorithm, the PCoA revealed the presence of two major groups albeit with some admixture among sub-groups (Appendix VI). Coordinate 1 and 2 explained 65.8% of the total genetic variation. The Bayesian-based clustering approach implemented in STRUCTURE also identified two groups according to the Δk approach (Figure 3.1b, c). Based on the estimated membership coefficient (Q), the South Asian, Western European and South American sub-groups ($Q > 0.70$) could be clustered together within G1, and the North American and Eastern European sub-groups ($Q > 0.70$) could be similarly clustered within G3. The North American/European sub-group (G2) was mostly an admixture of the other two major groups. Taken together, the NJ, PCoA and STRUCTURE analyses all agreed with respect to the distribution of the 407 flax accessions into two major groups. Additionally, the NJ and STRUCTURE analyses agreed in partitioning the collection into six sub-groups, with few differences among sub-groups. The high proportion of shared alleles revealed by the PCoA and STRUCTURE analyses was confirmed by the weak population structure as measured by the coefficient of population differentiation ($F_{ST} = 0.094$, $P < 0.001$) between G1 and G3. The level of differentiation between sub-groups ranged from 0.02 ($P < 0.001$, North America *vs* Eastern Europe) to 0.16 ($P < 0.001$, Eastern Europe *vs* South Asia) (Appendix VI).

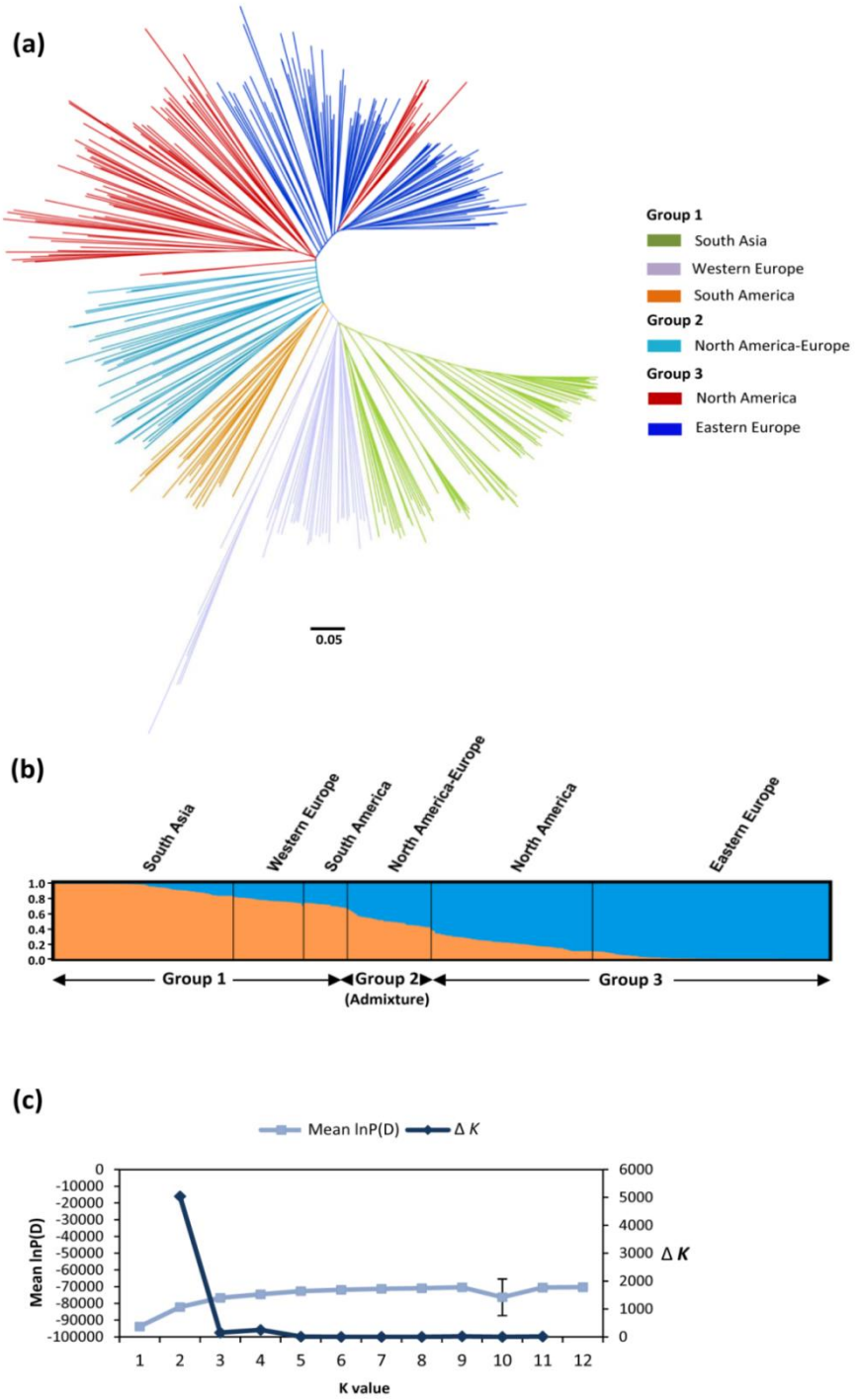


Fig. 3.1 Genetic relationships and population structure of the 407 flax accessions of the core collection. (**a**) Phylogenetic tree created using the Neighbor-joining (NJ) algorithm (Nei 1973) and information from 414 neutral SSRs. Colored clusters represent the sub-groups within major groups. The scale bar indicates the Nei (1973) minimum genetic distance. (**b**) Bayesian clustering (STRUCTURE, $K = 2$) of flax accessions. Sub-groups within Group 1 and Group 2 are distributed according to the clustering obtained by the NJ analysis. Accessions with a membership coefficient $Q < 0.7$ were classified as admixture Group 2. (**c**) Average log-likelihood values (mean lnP(D) ± SD for 10 iterations) and *ad-hoc* statistic ΔK (Evanno et al. 2005) for K values ranging from 1 to 12.

3.4.3 Molecular coancestry

Based on the alleles of the 448 microsatellites, the average molecular coancestry between any two flax accessions was 0.287 in the core collection as a whole. Approximately 70% of the pairwise coancestry estimates ranged from 0.1 to 0.3 (Figure 3.2a). The intra sub-group molecular coancestry ranged from 0.587 (Western Europe) to 0.713 (Eastern Europe). The pairwise molecular coancestry estimates ranged from 0.525 (North America *vs* Western Europe) to 0.633 (North America *vs* Eastern Europe) (Figure 3.2c). Overall, more than 80% of the pairwise molecular coancestry estimates in the core collection and sub-groups ranged from 0.114 to 0.350 and 0.525 to 0.601, respectively. The coancestry analysis indicated that most flax accessions had weak and moderate familial relatedness with the other accessions in the core collection and sub-groups respectively, which may be a reflection of the broad genetic diversity of the PGRC collection and the careful selection of accessions exercised while constructing the core collection.

3.4.4 Genetic diversity

In the core collection, the 414 neutral microsatellites retained detected 2202 alleles (N_a) (mean = 5.32/locus), out of which 1187 (54%) had a MAF < 0.05 and were considered rare alleles (R_a). The total unbiased gene diversity (UH_e) and the observed heterozygosity (H_o) were 0.427 and 0.023, respectively. Allelic richness (R_s) was estimated at 5.68, the inbreeding coefficient (F_{IS}) at 0.946 and the PIC value at 0.374. The genetic diversity parameters were also estimated for the major groups and sub-groups (Table 3.1). The parameters N_a, R_s, $\prod$, R_a and PIC in G1 were superior to those in G3, even though the population size of G1 was 25% smaller than G3. The parameters H_o and F_{IS} across the

core collection, the major groups and sub-groups are consistent with the predominantly self-pollinated nature of the species.

3.4.5 Linkage disequilibrium

To analyze LD variation, genetic distances for 293 microsatellites were available from the consensus linkage map of flax (Cloutier et al. 2012b). The average genome-wide distance between adjacent markers was 5.3 ± 2.4 cM.

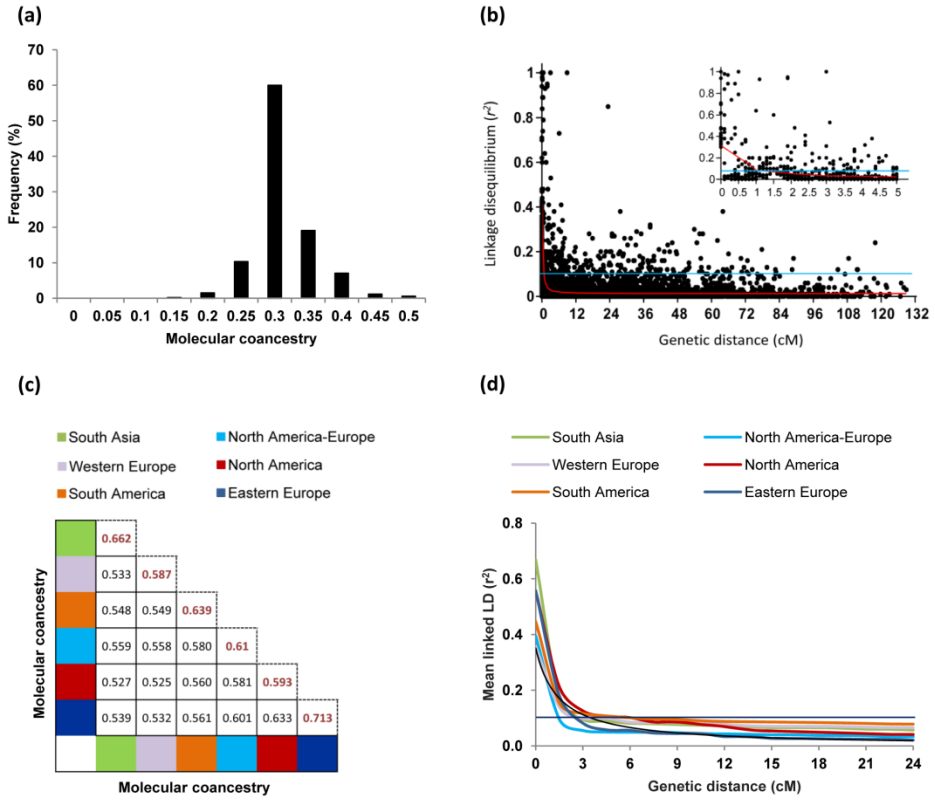


Fig. 3.2 Distribution of pairwise molecular coancestry estimates and linkage disequilibrium decay. (**a**) Global pairwise molecular coancestry estimates of the 407 flax accessions of the core collection. Only kinship values ranging from 0 to 0.5 are shown. (**b**) Scatter plot of LD decay (r^2) against the genetic distances (cM) for pairs of linked SSRs across the 15 linkage groups. The inner panel shows a detailed view of LD decay for markers located within 5 cM. The decay curves were plotted according to Breseghello and Sorells (2006). The blue line represents the threshold level of significance (r^2 = 0.1). The red line represents the average genome-wide LD of linked markers. (**c**) Pairwise molecular coancestry estimates (Caballero and Toro 2002) within each of the six sub-groups. The diagonal values correspond to the intra sub-group molecular coancestry. (**d**) Average genome-wide LD decay curve for linked markers within each of the six sub-groups.

Table 3.1 Genetic diversity parameters of the core collection, the two major groups (G1 and G3), the admixed group (G2) and their sub-groups.

Population	N[1]	UH_e[2]	H_o[3]	N_a[4]	R_s[5]	∏[6]	R_a[7]	F_{IS}[8]	Polymorphic loci (%)	PIC[9]
Core collection	407	0.427	0.023	2202	5.68	-	1187	0.946	100	0.374
Group 1	153	0.418	0.023	1978	4.37	547	925	0.944	99.8	0.361
South Asia	92	0.348	0.020	1510	2.85	116	542	0.931	95.9	0.305
Western Eur.	37	0.448	0.017	1608	3.44	246	418	0.961	97.1	0.393
South America	24	0.395	0.047	1135	2.70	27	186	0.878	91.3	0.332
Group 2										
N. Amer./Eur.	43	0.411	0.023	1341	2.91	32	324	0.933	96.4	0.352
Group 3	211	0.356	0.022	1613	3.44	183	683	0.933	99.1	0.332
North America	95	0.378	0.028	1362	2.69	73	424	0.932	98.6	0.334
Eastern Europe	116	0.300	0.020	1487	2.55	45	642	0.927	95.7	0.265

[1] Number of accessions
[2] Unbiased gene diversity
[3] Observed heterozygosity
[4] Number of alleles
[5] Allelic richness and [6] number of private alleles estimated on a sample of balanced size using the rarefaction method (Kalinowski 2005)
[7] Rare alleles < 5%
[8] Inbreeding coefficient
[9] Polymorphism information content

In the core collection, the average r^2 values for linked and unlinked markers were 0.036 and 0.023, respectively. The 95th percentile of r^2 distribution for unlinked markers was 0.09 and 10.81% of the loci pairs were in significant LD. The average genome-wide LD decayed to 0.1 within 1.5 cM (Figure 3.2b). LD values within sub-groups and major groups are presented in Table 3.2. The average r^2 values for linked and unlinked markers were higher in G1 than in G3 and the percentage of loci in significant LD was lower with 8.10% in G1 versus 12.22% in G3. Slower LD decay was observed within sub-groups (Figure 3.2d), ranging from 1.5 cM (North America-Europe) to 6.0 cM (South America), which could be attributed to the limited population size and narrow genetic diversity of some sub-groups as compared to the core collection. Regardless of the data set, i.e., core

collection or inferred groups, the average r^2 for linked markers remained higher than for unlinked markers, supporting physical linkage as the main determinant of LD in this core collection. The relatively rapid LD decay within the core collection suggested that high marker saturation will be required for effective AM. The slower LD decay within some of the sub-groups could be exploited for exploratory AM or coarse mapping.

3.4.6 Identification of non-neutral loci

The fiber and linseed groups made up of the 92 fiber accessions of the core collection and a random subset of 92 linseed accessions were subjected to bottleneck analysis (Cornuet and Luikart 1996). The mode-shift test showed the typical L-shaped distribution of allele frequencies in both groups (data not shown), expected at mutation drift-equilibrium when rare alleles are numerous, thus suggesting absence of a recent bottleneck. The sign test, however, indicated a heterozygosity excess (bottleneck) in the fiber group ($P < 0.01$) but not in the linseed group ($P = 0.346$). The population structure analysis showed a sharp peak of Δk at $K = 2$ largely corresponding to the fiber and linseed types (Appendix VII). Thus, no hierarchical population structure was detected and the two original groups (fiber and linseed) were adopted for posterior analyses.

Distortion from neutral expectations was detected at 41, 13, 14 and 26 loci with EW, ln RH, LOSITAN and Arlequin, respectively (data not shown). A total of 9 loci (mean $F_{ST} = 0.16$) distributed across 7 linkage groups were significant in at least two of the four tests and were considered true outliers (Table 3.3). LD between these and their adjacent loci ranged from 0 to 0.10 in the fiber group and from 0 to 1.0 in the linseed group.

The physical to genetic distance between c306-s98_Lu765Bb and its closest locus c306-s98_Lu3063 was estimated at 364 kb/cM. Considering an LD of 1 between them and a genetic distance estimate of 1.3 cM, we investigated a physical interval of 474 kb.

Table 3.2 Linkage disequilibrium in the core collection, the two major groups (G1 and G3), the admixed group (G2) and their six sub-groups.

Population	Mean linked LD (r^2)	Mean unlinked LD (r^2)	95th percentile unlinked LD	Loci pairs in significant LD (%)
Core collection	0.036	0.023	0.09	10.81
Group 1	0.047	0.035	0.14	8.10
South Asia	0.070	0.056	0.22	8.82
Western Eur.	0.072	0.067	0.26	6.08
South America	0.084	0.067	0.25	8.75
Group 2				
North Amer./Eur.	0.040	0.032	0.12	8.08
Group 3	0.037	0.019	0.08	12.22
North America	0.061	0.030	0.11	15.68
Eastern Europe	0.036	0.020	0.08	10.86

A total of 98 genes were predicted in this interval, of which 59 showed significant similarities with high quality annotations of UniProtKB protein database (Appendix VIII). The physical to genetic distance between c16-s156_Lu373 and c16-s156_Lu139 was estimated at 178 kb/cM with a moderate LD ($r^2 = 0.22$). Consequently, the hitchhiking effects may not extend across the total genetic distance of 1.9 cM. We estimated the LD decay in both flax type groups to calculate an average genetic distance at which LD was strong (Appendix VII). LD decayed to 0.4 within ≈ 0.2 cM, equivalent to 36 kb in which only five highly similar genes were predicted. The number of predicted genes from different scaffolds with weak LD ($r^2 < 0.2$) having significant similarities to annotated proteins ranged from 3 (c175-s86_Lu2824, c206-s208_Lu128 and c441-s225_Lu3189) to 5 (c36-s291_Lu176 and c108-s305_Lu595) (Appendix VIII).

GO annotations could be assigned to ~60% of the predicted genes which are expressed in flax based on EST and protein evidence. Mapping of predicted proteins from 86 candidate genes to the UniProtKB database yielded 1,035 GO annotations as a result of multiple associations of individual proteins with multiple functions, processes or components (Ashburner et al. 2000) (Appendix VIII). The top four GO categories for molecular function were 'binding' (21.9%), 'catalytic activity' (14.4%), 'nucleotide binding' (10.6%) and 'hydrolase activity' (10%) (Appendix IX). Similarly, functional characterization of proteins associated with candidate genes at non-neutral loci indicated unknown broad ‘biological processes' (19.1%), followed by 'cellular processes' (16%), 'metabolic processes' (11.1%) and 'response to stress' (7.6%) (Appendix IX). The candidate gene products were localized to membrane (11.8%) and intracellular locations (9.2%) (Appendix IX). Approximately 4.6% of the predicted proteins were localized to the cell wall. Key genes associated with cell-wall biogenesis/modification, xylem identity, auxin regulation and fatty acid biosynthesis were identified among our candidate genes potentially affected by divergent selection in flax (Appendix VIII).

3.5 Discussion

To contribute to long term sustainability of flax production and diversification, the germplasm stored in PGRC has comprehensively been characterized for morphologic, phenologic and agronomic characters (Diederichsen and Fu 2008). This valuable phenotypic information enabled the construction of a flax core collection of 407 accessions to further flax genetic studies and improvement. Here, we report on the genetic characterization of the core collection based on 448 microsatellite loci which represents one of the largest flax genetic studies published to date (Månsby et al. 2000;

Spielmeyer et al. 1998; Everaert et al. 2001; Fu et al. 2003; Wiesnerová and Wiesner 2004; Fu 2005, 2011; Diederichsen and Fu 2006; Cloutier et al. 2009, 2012a; Rajwade et al. 2010; Uysal et al. 2010; Rachinskaya et al. 2011; Soto-Cerda et al. 2012).

3.5.1 Genetic relationships and population structure

Understanding the genetic relationships and structure of core collections is critical to control false positives in AM (Myles et al. 2009). The NJ tree grouped the 407 flax accessions mainly but not exclusively according to geographical origin. The presence of accessions from countries out of the geographical clusters could be explained by the fact that the passport data may be occasionally weak where the donor country is considered the country of origin. As a consequence, the names of the sub-groups were assigned according to the geographic origin of the majority of the accessions within them.

The South Asian sub-group of G1 was the most genetically distinct. Fu (2005) reported similar results in 2727 flax accessions assessed with 149 RAPD markers. However, in his study, the Indian subcontinent and Central Asia were considered related groups rather than a unified cluster. Differences in the marker systems and extent of the genome coverage (414 mapped microsatellite *vs*. 149 RAPD markers) could explain the resolution differences between studies. The active exchanges of flax germplasm between France, Germany, the United Kingdom and Hungary provide support for the Western European grouping (Maggioni et al. 2002). The genetic relationships among G1 accessions were also supported by a weak population differentiation among sub-groups (F_{ST} = 0.05 - 0.11, Appendix VI). Within G3, the North American sub-group reflects historical germplasm exchange between the U.S.A. and Canada (Fu et al. 2003).

Table 3.3 Outlier analysis for divergent selection between fiber and linseed types.

Locus	Linkage group	Outlier analysis				Highest LD[1]		F_{ST}
		Ewens-Watterson	Ln RH	LOSITAN	Hierarchical	Fiber	Oil	
c206-s208_Lu128	2	n.s.	n.s.	***	**	0.03 (6.2)	0.03 (6.2)	0.43
c475-s917_Lu2021a	2	*	**	***	n.s.	0.01 (1.5)	0.03 (8.3)	0.04
c16-s156_Lu373	3	**	*	n.s.	n.s.	0.10 (1.9)	0.22 (1.9)	0.16
c36-s291_Lu176	5	**	n.s.	n.s.	*	0.02 (3.3)	0.06 (3.3)	0.27
c108-s305_Lu595	8	**	**	***	n.s.	0.00 (0.0)	0.00 (0.0)	0.10
c441-s225_Lu3189	8	*	n.s.	n.s.	*	0.05 (2.7)	0.31 (2.7)	0.32
c175-s86_Lu2824	9	*	**	***	n.s.	0.01 (4.1)	0.02 (6.9)	0.02
c306-s98_Lu765Bb	12	**	**	***	n.s.	0.09 (1.3)	1.00 (1.3)	0.07
c226-s280_Lu637	15	**	*	***	n.s.	0.02 (3.7)	0.13 (3.7)	0.01

* $P < 0.05$, ** $P < 0.01$, *** false discovery rate (FDR) < 0.05, n.s. not statistically significant

[1] highest LD between the candidate locus and one of the two adjacent loci. The genetic distance (cM) at which the highest LD extent was observed is indicated in brackets

The Eastern European sub-group contained most of the fiber flax accessions from the Netherlands and the former Soviet Union but it also included linseed accessions that were not intermixed. They were separated by a small group of U.S.A. accessions clustered within this sub-group. The U.S.A. accessions were mostly fiber type. Similar results observed in the population structure analyses and the lowest F_{ST} (0.02) between sub-groups (Appendix VI) could explain the interstitial presence of the U.S.A. accessions. The two major groups supported by our combined approach showed weak population subdivision in support of the breadth of the genetic diversity captured in this collection, making it ideal for AM (Flint-Garcia et al. 2003).

3.5.2 Molecular coancestry

Strong population structure, familial relatedness, or both, may be significant in a core collection and would negatively impact AM. Yu et al. (2006) developed a mixed linear model (MLM) which incorporates the pairwise kinship (*K* matrix) to correct for relatedness. Spurious associations cannot be controlled completely by population structure (*Q* matrix) (Yu et al. 2006; Myles et al. 2009). Models incorporating a *K* matrix are generally superior in controlling the rate of false positives while maintaining statistical power as compared to those using only a *Q* matrix (Yu et al. 2006).

In self-pollinated crops or inbred lines, coancestry estimates tend to be higher than in outcrossing species because the high heterozygosity reduces the probability that two alleles observed at a locus are identical by state (Bernardo et al. 2000). In our core collection, approximately 80% of the pairwise coancestry estimates ranged from 0.1 to 0.3, indicating that most of the lines had weak relatedness (Figure 3.2a). We anticipate

that with the weak population structure and relatedness of the core collection, a MLM correcting for *K* should provide sufficient statistical power to control most of the false positive associations in future AM studies (Yu et al. 2006).

3.5.3 Genetic diversity

A suitable core collection for AM should encompass as much phenotypic and molecular diversity as can be reliably measured in a given environment (Flint-Garcia et al. 2003; Myles et al. 2009). An average of 5.32 alleles per locus over 414 microsatellites was observed in our core collection. This value is higher than the range previously reported (2.72 – 3.46) (Roose-Amsaleg et al. 2006; Deng et al. 2010; Rachinskaya et al. 2011; Soto-Cerda et al. 2012). This allelic diversity even exceeded that of a diverse sample of *L. usitatissimum* L. subsp. *angustifolium* (Huds.) Thell., (wild progenitor) and *L. usitatissimum* L. subsp. *usitatissimum* (4.62) (Fu 2011). This may be the result of the number of genotypes analyzed (407), the choice of the germplasm, the number of microsatellite loci (414 neutral out of 448) and the microsatellite repeat type and length (Vigouroux et al. 2002; Cloutier et al. 2009).

A higher number of private alleles were observed in G1 as compared to G3 (Table 3.1). The Western European sub-group was particularly rich in private alleles with 246. Novel genetic variations, not previously sampled or utilized in modern flax breeding programs, may be present in this sub-group, offering unique alleles for broadening the diversity of flax gene pools. This is contrary to previous studies that have reported generally low genetic diversity of flax germplasm (Fu et al. 2003; Diederichsen and Fu 2006; Cloutier et al. 2009; Fu 2011; Smỳkal et al. 2011; Soto-Cerda et al. 2012).

Although 85% of the accessions of our core collection are cultivars and breeding materials, the collection possesses abundant genetic diversity, an advantageous attribute for dissecting the genetic basis of QTL for immediate application in flax breeding (Flint-Garcia et al. 2003; Yu and Buckler 2006).

3.5.4 Linkage disequilibrium

Low LD demands the use of dense marker sets resulting in tight linkage between markers and QTL, an advantageous criterion for breeding applications because the predictive ability of a marker will be robust through generations (Flint-Garcia et al. 2003). The average r^2 of the entire core collection was 0.036 and the average genome-wide LD decayed within 1.5 cM (Figure 3.2b). In self-pollinated species where recombination is less effective than in outcrossing species LD declines more slowly (Flint-Garcia et al. 2003). Nonetheless, the germplasm that makes up the collection plays a key role in LD variation because the extent of LD is influenced by the level of genetic variation captured by the target population. For example, in wild barley (*Hordeum vulgare* ssp. *spontaneum*), despite its high rate of self-fertilization (~98%), LD decayed within 2 kb, a value similar to that observed in maize, an outcrossing species (Morrell et al. 2005). The low LD of this core collection dictates the need for a higher marker saturation to provide superior mapping resolution and QTL detection power by AM (Xiao et al. 2012) as compared to using biparental linkage maps. Alternatively, selection of sub-groups with low F_{ST} and higher but similar levels of LD would require a reduced number of individuals and markers for exploratory AM.

The percentage of loci pairs in significant LD was fairly similar in each sub-group except for the North American and Eastern European sub-groups which registered the highest values, possibly reflecting their more intensive artificial selection and narrow germplasm (Fu et al. 2003). Although our core collection did not behave as an unstructured large population, our combined analyses of population structure showed that G1 and G3 were weakly differentiated, representing two ancestral populations that minimize differences in LD and potentially the amount of spurious associations (Figures 3.1a, b). Thus, the results of our LD characterization within diverse genetic groups offer the versatility to perform cost-effective AM studies in flax by providing the fundamental characterization of the collection demonstrating its usefulness for AM.

3.5.5 Identification of non-neutral loci

Flax is one of the few domesticated plants that has been subjected to disruptive selection (Cullis 2011). North America almost exclusively grows linseed and, up until recently, the stems were considered more problematic than beneficial because of their slow field biodegradation. However, the use of short fibers has received increased attention in North America in the last few years because of the interest in extracting value from the stem of linseed varieties (Diederichsen and Ulrich 2009). Stem fiber content does not seem associated with qualitative or quantitative plant characters in flax germplasm (Diederichsen and Ulrich 2009) indicating that there are no major biological restrictions for pyramiding agronomic and seed quality traits with high fiber content.

Crops have been subjected to strong selective pressure directed at genes controlling traits of agronomic importance during their domestication and subsequent

episodes of selective breeding (Vigouroux et al. 2002). Under positive selection, favourable alleles will increase in frequency until fixation. As an effect of genetic hitchhiking, loci closely linked to beneficial alleles might present distortions from neutral expectations. Genome scans have allowed the identification of candidate loci involved in domestication and breeding traits in several crops (Vigouroux et al. 2002; Casa et al. 2005) and domesticated animals (Flori et al. 2009; Schwarzenbacher et al. 2012). However, population structure and bottlenecks can mimic the effect of selection and create false positives. The combination of several methods based on different assumptions can reduce false positives (Shimada et al. 2011).

We applied four different tests of neutrality to identify the genomic regions that deviate from neutral expectations potentially associated with fiber and linseed divergent selection. Collectively, 86 candidate genes were identified at nine loci (Appendix VIII). Among our candidate genes, we found a β-tubulin involved in cell morphogenesis and elongation of fiber in cotton (He et al. 2008), a glucan endo-1,3-β-glucosidase associated with cell wall biogenesis/degradation in flax (Roach and Deyholos 2008), a chitinase involved in polysaccharide degradation (Roach and Deyholos 2008), a MYB transcription factor that influences cellulose microfibril angle in Eucalyptus (Sexton et al. 2011) and a class III HD-Zip protein 4 (HB4) involved in xylem identity in flax (Fenart et al. 2010) (Appendix VIII). Candidate genes such as pyruvate dehydrogenase E1 and fatty acid alpha-hydroxylase involved in fatty acid biosynthetic processes were also identified (Appendix VIII). However, β-galactosidase and cellulose synthase, two key enzymes for cell-wall modification and cellulose synthesis in flax (Roach and Deyholos 2008; Fenart et al. 2010) were not present at any of the nine loci. Previously identified genes in flax

microarray analyses of hypocotyl and phloem fiber development (Roach and Deyholos 2008) and differentially expressed genes between flax inner and outer stem tissues (Fenart et al. 2010) were found among our candidate genes (Appendix VIII).

Although preliminary, our scans provided the first insights of non-neutral loci potentially affected by divergent selection in flax. Candidate genes, especially those previously reported (Roach and Deyholos 2008; Fenart et al. 2010), will require further investigation and validation. To enhance the probability of identifying additional candidate loci, a high density of markers would be desirable. Currently, next generation sequencing technology enables the re-sequencing of a high number of accessions at a reasonable price. Thus, high quality and dense single nucleotide polymorphism (SNP) markers promise to provide comprehensive genome coverage for the identification of non-neutral genomic regions in flax (Schwarzenbacher et al. 2012). Such genomic tools for flax genetic studies are being developed and more comprehensive genomic scans will be possible in the near future.

3.6 Conclusion

In this study, high levels of genetic diversity were revealed as compared to previous flax genetic studies. The weak population structure and relatedness and relatively fast LD decay indicate the suitability of this flax core collection for AM. The peculiar divergent breeding applied in the development of fiber and linseed flax varieties provides a unique opportunity to understand how human needs have sculpted the flax genome during domestication and improvement, and how these divergent genomic regions could be deployed in breeding for flax as a dual purpose crop.

ASSOCIATION MAPPING OF SEED QUALITY TRAITS USING THE CANADIAN FLAX (*LINUM USITATISSIMUM* L.) CORE COLLECTION

Braulio J. Soto-Cerda[1,2] · Scott Duguid[3] · Helen Booker[4] · Gordon Rowland[4] · Axel Diederichsen[5] · Sylvie Cloutier[1,2]

[1]University of Manitoba, Department of Plant Science, 66 Dafoe Road, Winnipeg, MB, R3T 2N2, Canada

[2]Cereal Research Centre, Agriculture and Agri-Food Canada, 195 Dafoe Rd, Winnipeg, MB, R3T 2M9, Canada

[3]Morden Research Station, Agriculture and Agri-Food Canada, Route 100, Morden, MB, R6M 1Y5, Canada

[4]University of Saskatchewan, Crop Development Centre, College of Agriculture and Bioresources, 51 Campus Drive, Saskatoon, SK, S7N 5A8, Canada

[5]Plant Gene Resources of Canada, Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK, S7N 0X2, Canada

The author Braulio J. Soto-Cerda carried out the analyses, interpretation of data and co-wrote the manuscript. The major supervisor Dr. Sylvie Cloutier designed the study, generated the data, and co-wrote the manuscript. Dr. Scott Duguid conducted the field trials in Manitoba and carried out the chemical analyses of the seed quality traits. Drs. Helen Booker and Gordon Rowland conducted the field trials in Saskatchewan. Dr. Axel Diederichsen characterized and developed the flax core collection.

The manuscript was submitted to **Theoretical and Applied Genetics** and it is in press.

4.0 ASSOCIATION MAPPING OF SEED QUALITY TRAITS USING THE CANADIAN FLAX (*LINUM USITATISSIMUM* L.) CORE COLLECTION

4.1 Abstract

Linseed oil is valued for its food and non-food applications. Modifying its oil content and fatty acid (FA) profile to meet market needs in a timely manner requires clear understanding of their quantitative trait loci (QTL) architectures, which have received little attention to date. Association mapping is an efficient approach to identify QTL in germplasm collections. In this study, we explored the quantitative nature of seed quality traits including oil content (OIL), palmitic acid (PAL), stearic acid (STE), oleic acid (OLE), linoleic acid (LIO), linolenic acid (LIN) and iodine value (IOD) in a flax core collection of 390 accessions assayed with 460 microsatellite markers. The core collection was grown in a modified augmented design at two locations over three years and phenotypic data for all seven traits were obtained from all six environments. Significant phenotypic diversity and moderate to high heritability for each trait (0.73-0.99) were observed. Most of the candidate QTL were stable as revealed by multivariate analyses. Nine candidate QTL were identified, varying from one for OIL to three for LIO and LIN. Candidate QTL for LIO and LIN co-localized with QTL previously identified in bi-parental populations and some mapped nearby genes known to be involved in the FA biosynthesis pathway. Fifty-eight percent of the QTL alleles were absent (private) in the Canadian cultivars suggesting that the core collection possesses QTL alleles potentially useful to improve seed quality traits. The candidate QTL identified herein will establish the foundation for future marker-assisted breeding in linseed.

4.2 Introduction

Oil crops have gained in importance worldwide over the past 20 years as indicated by the increase in total harvested area from 189.3 million hectares in 1992 to 272.7 million hectares in 2011 (FAOSTAT 2013). This increase hinges partly on the versatility of their fatty acid profiles which play a significant role in the nutritional properties and the end-use functionality of oil crops. In this regard, linseed (*Linum usitatissimum* L.), with its high content of alpha linolenic acid, is unique. With ~23% of the world production, Canada is the world's largest linseed producer and exporter followed by China and the Russian Federation (FAOSAT 2013).

Linseed is an annual, self-pollinated species with a genome size of ~ 370 Mb (Ragupathy et al. 2011). Domesticated in the Near East 9,000 years ago (Harris 1997), linseed is considered the oldest oilseed in the world. Its seed oil (~35-50%) is composed of five main fatty acids (FAs): palmitic (PAL; C16:0, ~6%), stearic (STE; C18:0, ~2.5%), oleic (OLE; C18:1, ~19%), linoleic (LIO; C18:2, ~13%) and linolenic (LIN; C18:3, ~55%) (Westcott and Muir 2003; Diederichsen et al. 2013). The high percentage of LIN distinguishes it from other oilseeds in the industrial, human food and animal feed markets. Its oxidative instability, ensuing in a soft and flexible film, and the absence of volatile organic compounds (formaldehyde, aldehydes and benzene), resulting in reduced environmental hazards (Green et al. 2008), makes linseed oil valuable in industry for paints, linoleum flooring, inks and varnishes (Cullis 2007). In addition, LIN is the precursor of the long chain polyunsaturated fatty acids eicosapentaenoic acid (EPA), docosapentaenoic acid (DPA) and docosahexaenoic acid (DHA) which are synthesized in the human body and recognized for their health benefits (Simopoulos 2000).

Linseed breeders have focused mainly on maintaining the high LIN content while PAL, STE, OLE and LIO which correlate negatively with LIN tend to be selected against (Cullis 2007). High LIN (>65%) germplasm is available (Friedt et al. 1995; Kenaschuk 2005) but agronomic improvement of many of these sources is needed to achieve adaptability. The first high LIN linseed cultivar NuLin^{TM} 50 was registered in Canada by Viterra (http://www.viterra.ca). Altered FA profiles in linseed, for example low LIN (2-4%) and high LIO (>50%) obtained by mutation breeding, has proven effective in improving the oxidative stability and suitability of linseed oil for a variety of food uses (Green et al. 2008). Green and Marshall (1984) developed linseed lines with reduced LIN content (< 29%) using ethyl methanesulphonate (EMS)-mediated mutagenesis. Further reduction in LIN content to ~2% was later achieved (Green 1986; Rowland 1991). Fatty acid desaturase 3 genes *lufad3a* and *lufad3b* had point mutations causing premature stop codons in one of the characterized EMS mutant lines resulting in non-functional FAD3 enzymatic activity (Vrinten et al. 2005). Additional variations in FA composition, including lines with elevated OLE and PAL content, have also been developed (Green, unpublished data; Rowland and Bhatty 1990).

Various aspects of the genetic control of storage oil biosynthesis in linseed have been studied (Green 1986; Fofana et al. 2004; SØrensen et al. 2005; Vrinten et al. 2005; Khadake et al. 2009; Banik et al. 2011) and new genes such as *LuFAD2-2* (Khadake et al. 2009) and *fad3c* (Banik et al. 2011) encoding FA desaturases have been cloned, broadening the options for modifying linseed FA profiles for new end-uses. Generally, oilseed breeding is a more complex undertaking than the breeding of cereals or legumes, as many oilseeds such as soybean, rapeseed, sunflower and linseed have the potential to

be dual- or multi-purpose crops, which requires the simultaneous manipulation of quality and agronomic traits (Vollmann and Rajcan 2009). Conventional breeding has been conducted in linseed for over a century and has been particularly successful in adapting crop phenology to regional growing seasons as well as providing yield stability across environments (Green et al. 2008). However, the phenotypic selection of quantitative traits, such as oil content and FA composition is complicated by environmental effects (Cloutier et al. 2011) that significantly reduce breeding gain. In Canada, oil content can vary up to 15% (range 35-50%) in individual farm samples (Duguid 2009) and the percentage of LIN can be as much as 5% higher in cool environments (Fofana et al. 2006).

Consumer awareness of oil quality is becoming an increasingly important variable that conditions shifts in the food ingredient selection process, thereby creating new market opportunities (Wilson 2012). Acceleration of breeding cycles could translate into the edge necessary to respond to these new market demands in a timely fashion.
The use of marker-assisted selection (MAS) for oil content and FA composition can improve the efficiency of traditional linseed breeding. However, MAS requires the development of genomic tools such as molecular markers and linkage maps (Cloutier et al. 2009, 2011, 2012a, b). These tools have been recently developed in linseed, establishing the foundation for the application of MAS (Roose-Amsaleg et al. 2006; Cloutier et al. 2009, 2011, 2012a, b; Deng et al. 2010, 2011; Ragupathy et al. 2011; Soto-Cerda et al. 2011a, b; Venglat et al. 2011; Kumar et al. 2012; Wang et al. 2012a).

Quantitative trait loci (QTL) mapping based on bi-parental crosses has been the most applied approach to map QTL associated with oil content and FA in crops such as

rapeseed (Zhao et al. 2005; Hu et al. 2006; Qiu et al. 2006; Smooker et al. 2011); maize (Goldman et al. 1994; Wassom et al. 2008; Yang et al. 2010) and soybean (Chung et al. 2003; Bachlava et al. 2009; Qi et al. 2011; Xie et al. 2012). In linseed, however, only one QTL study related to oil content and FA composition has been carried out, positioning QTL for iodine value (IOD), PAL, LIO and LIN (Cloutier et al. 2011). While QTL mapping has been very successful in detecting QTL, the bi-parental nature of the populations often resulted in large confidence intervals for the QTL positions which, combined with a limited number of alleles at each locus, hindered their applications in MAS (Gupta et al. 2005; Ersoz et al. 2009; Myles et al. 2009).

Association mapping (AM) or linkage disequilibrium (LD) mapping has emerged as a complementary approach to QTL mapping (Myles et al. 2009). Its power relies on the utilization of a large population of individuals with a higher level of allelic diversity that improves the probability of QTL detection and the mapping resolution (Ersoz et al. 2009). AM has been useful in dissecting the complex genetic architecture of oil content and FA composition in oil crops such as rapeseed (Honsdorf et al. 2010; Zou et al. 2010); peanut (Wang et al. 2011); soybean (Li et al. 2011d) and maize (Cook et al. 2012; Li et al. 2013). These AM studies not only validated previous results from QTL mapping showing the FA biosynthesis pathway similarity among oil crops, but also identified new QTL and candidate genes useful for improving oil content and quality.

In our previous study, we characterized the flax core collection of 407 accessions assembled from the Canadian flax world collection preserved by Plant Gene Resources of Canada (Diederichsen et al. 2013), and showed its abundant genetic diversity, weak population structure and familial relatedness, and a relatively fast LD decay, all positive

attributes for AM studies (Soto-Cerda et al. 2013). In the present study, we conducted AM for oil content and FA composition traits on 390 accessions aiming to identify QTL underlying these seed quality traits, which could be used for accelerating linseed breeding through MAS and for identifying germplasm with desirable characteristics.

4.3 Materials and methods

4.3.1 Plant material, genotyping and field experiments

A core collection of 407 *L. usitatissimum* accessions assembled from the Canadian World collection of flax (~3,500 accessions) (Diederichsen et al. 2013) was genotyped with 460 microsatellite (SSR) markers (Roose-Amsaleg et al. 2006; Cloutier et al. 2009, 2012a; Deng et al. 2010, 2011) distributed across the 15 linkage groups of flax (Cloutier et al. 2012b). The amplification products were resolved on an ABI 3130xl Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). Output files were analyzed by GeneScan (Applied Biosystems) and subsequently imported into Genographer. Fragment sizes were estimated using GeneScan ROX-500 (Applied Biosystems) and MapMarker® 1000 (BioVentures Inc., Murfreesboro, TN) internal size standards. The genotype of each locus was encoded based on its allele size in bp or as a null allele for dominant markers.

The flax core collection was assessed with 259 mapped neutral SSR loci which indicated that all accessions were organized into two major groups (G1 and G3) and one admixed group (G2) with a weak population structure (F_{ST} = 0.09) (Soto-Cerda et al. 2013). G1 included 90% of the fiber flax accessions mostly from Western Europe and linseed accessions from South Asia and South America while G3 included accessions

from North America and Eastern Europe and was mostly oil type. A relatively fast genome-wide LD decay of ~1 cM (r^2 = 0.1) was estimated (Soto-Cerda et al. 2013).

Phenotypic data was collected from 390 accessions including 381 accessions selected by Diederichsen et al. (2013) and nine accessions of relevance to recent Canadian flax breeding programs. The 390 accessions were evaluated during three years (2009, 2010 and 2011) in Morden, Manitoba (MB) and at the Kernen Farm located near Saskatoon, Saskatchewan (SK), Canada, which represent two mega-environments where most of the linseed is produced in Western Canada (http://www.canadagrainscouncil.ca/). A type 2 modified augmented design (MAD) (Lin and Poushinsky 1985) was used to phenotype oil content and FA composition traits. Main plots (2 m long, 2 m wide with 20 cm row spacing) were arranged in grids of 10 rows and 10 columns. Each main plot was divided into five paralleled subplots of two rows each with a plot control (CDC Bethune replicated 100 times) located in the centre. Additional subplot controls (Hanley and Macbeth) were assigned to five randomly selected main plots. The 4-m^2 plots were harvested, threshed and cleaned. Seeds of each plot were subsampled for oil content and FA composition analyses.

4.3.2 Oil content and FA analyses

OIL was measured by nuclear magnetic resonance calibrated against the FOSFA (Federation of Oils, Seeds and Fats Associations Limited) extraction method. Methyl esters of FA were prepared according to the American Oil Chemists' Society (AOCS) (http://www.aocs.org/Methods/index.) Official Method Ce 2-66 (09) and FA composition was determined by capillary gas chromatography (GC), following the AOCS Official

Method Ce 1e-91. IOD, a measure of the saturation level of lipids, was calculated from the GC-derived FA composition, following the AOCS Method Cd 1c-85.

4.3.3 Statistical analysis

Adjusted data was obtained for each trait as previously described based on the MAD (You et al. 2013). Normality of the adjusted data was tested using the Shapiro-Wilk test (Shapiro and Wilk 1965) and normal probability plots. The adjusted phenotypic values were used to estimate the variance components to determine the effect of year, location, genotype and their interactions on oil content and FA composition using the GLM procedure in SAS 9.1 (SAS Institute 2004) as described in You et al. (2013). As a measurement of the repeatability of the field trials across years within locations, broad sense heritability (*H*) on an entry mean basis for each seed quality trait was estimated as follows: $H = \sigma^2_G / [\sigma^2_G + (\sigma^2_{GE} / e) + (\sigma^2_E / e\, r)]$ where σ^2_G, σ^2_{GE}, σ^2_E, *e* and *r* correspond to the genetic variance, the genetic by environment interaction variance, the residual variance, the number of environments and the replications per environment, respectively. Pearson's correlation coefficients ($P < 0.001$) were used to express the relationships between seed quality traits.

4.3.4 Linkage disequilibrium

A LD heat map was constructed using six linkage groups (LGs) and 158 SSR loci (mean = 1 locus / 3.5 cM). The six LGs were selected based on their marker density and differences in size from the consensus linkage map of flax (Cloutier et al. 2012b). The heat map was produced with GGT 2.0 (van Berloo 2008) based on pair-wise r^2 estimates

for all marker pairs with minor allele frequency (MAF) > 0.05 (Breseghello and Sorrells 2006). Allelic frequencies were calculated in GENALEX v.6.41 (Peakall and Smouse 2006) and MAF < 0.05 were set to "U" (missing data) and excluded from the LD analysis. This heat map verified the relationships between genomic regions harboring significant markers and large blocks of LD. The 95th percentile of the distribution of unlinked markers $r^2 = 0.09$ (Soto-Cerda et al. 2013) was used to set the statistical r^2 value to determine LD that resulted from physical linkage (Breseghello and Sorrells 2006). Markers on different linkage groups were considered unlinked.

4.3.5 Association mapping

The adjusted phenotypic values of the seed quality traits were used for AM. Five AM models were tested in TASSEL 2.1 (Bradbury et al. 2007) including two GLMs and three mixed linear models (MLMs). The first GLM incorporated the *Q* matrix as the fixed covariate while the second used PCA (Price et al. 2006). The first MLM incorporated the kinship matrix (*K*) (Yu et al. 2006) as a random effect only, while the second and third used in addition the *Q* matrix and PCA as fixed covariates, respectively. The *Q* matrix was estimated using 259 mapped neutral SSRs (Soto-Cerda et al. 2013). The PCA matrix calculated in TASSEL 2.1 retained the first three components explaining 27% of the variation. The *K* matrix was constructed on the basis of 448 SSRs using SPAGeDi (Hardy and Vekemans 2002). All negative values between individuals were set to zero (Yu et al. 2006). The most suitable AM model was selected using cumulative probability-probability (P-P) plots which indicate the extent to which the analysis produced more

significant results than expected by chance. For the AM analysis, only MAF > 0.05 were retained (Breseghello and Sorrells 2006).

AM analyses for the seed quality traits were carried out for each year and location independently. Correction for multiple testing was performed using the *q*FDR value, which is an extension of the false discovery rate (FDR) method (Benjamini and Hochberg 1995). The *q* values were calculated with the QVALUE R package using the smoother method (Storey and Tibshirani 2003). Markers with *q*FDR < 0.01 in at least two years were considered significant within location. Further, markers with *q*FDR < 0.01 in at least four of the six environments were considered consistent associations. For markers significantly associated with a trait, a GLM with all fixed-effect terms was used to estimate the amount of phenotypic variation explained by each marker (R^2). Allelic effects of the significant marker loci were calculated as the difference between the average phenotypic values of the homozygous alleles with MAF > 0.05. The significant differences between the allele means were estimated by the Kruskal-Wallis non-parametric test (Kruskal and Wallis 1952) and visualized as box plots.

Candidate QTL were delineated using the estimated background LD (95^{th} percentile) for unlinked markers $r^2 = 0.09$ (Soto-Cerda et al. 2013) as suggested by Breseghello and Sorrells (2006). Thus, associated markers were considered linked and part of the same candidate QTL if they showed $r^2 > 0.09$. Since markers in the same QTL were closely linked and in significant LD, the amount of phenotypic effect explained by the candidate QTL was estimated using the marker within the QTL with the highest *P* value as described above for the significant markers.

4.3.6 QTL effect and stability

The QTL/marker effects were calculated as described above for the allelic effects. The stability of a candidate QTL and associated markers was estimated using the additive main effect and multiplicative interaction (AMMI) model (Zobel et al. 1988; Gauch 1992) in GenStat 14 (VSN International, 2011). Candidate QTL/markers with a first interaction principal component (IPCA1) near zero are more stable, while those QTL/markers with IPCA1 either positive or negative are more unstable. The AMMI's stability values (ASV) (Purchase 1997) were also calculated using the following formula: $\text{ASV} = \sqrt{\frac{\text{SSIPCA1}}{\text{SSIPCA2}}(\text{IPCA1})^2 + (\text{IPCA2})^2}$, where SSIPCA1 is the sum of squares interaction of the first principal component (PC) analysis and SSIPCA2 is the sum of squares interaction of the second PC analysis. The smaller the ASV value, the more stable the candidate QTL/markers are across environments. The stability of QTL/markers based on their IPCA1 was defined as follows: 0 to ± 0.5 highly stable; ± 0.51 to ± 1 stable; ± 1.01 to ± 1.5 moderately stable and higher than ± 1.51 unstable. The stability of QTL/markers based on their ASV values was defined as follows: 0 to 0.5 highly stable; 0.51 to 1 stable; 1.01 to 1.5 moderately stable and higher than 1.51 unstable. The QTL/marker effects estimated were decomposed into PCs via singular value decomposition and the first two PCs were plotted for both QTL/markers and environments to form a QTL main effect and QTL by environment interaction (QQE) biplot (Yan and Tinker 2005) using GenStat 14 (VSN International, 2011).

4.3.7 Frequency of QTL/marker allele in the flax core collection and Canadian cultivars

QTL/marker alleles were defined as alleles of the marker with the largest P-value from a QTL or alleles of a significantly associated marker not part of a candidate QTL. With the aim of identifying new potentially favourable QTL/marker alleles absent in linseed Canadian cultivars, the observed number of alleles, the number of private alleles and the allelic richness were contrasted for the 30 linseed Canadian cultivars (Appendix X) present in the flax core collection with the remaining 377 of diverse origins (Diederichsen et al. 2013; Soto-Cerda et al. 2013). In addition to the QTL, stable associated markers, not part of a QTL but that explained at least 1% of the phenotypic variation were also included. The number of private QTL/marker alleles and QTL/marker allelic richness were corrected for sample size differences and estimated using the rarefaction method implemented in HP-RARE v.1.2 (Kalinowski 2005). This analysis included all alleles, even the rare ones (MAF < 0.05). The frequencies of the most favourable QTL/marker alleles were estimated in GENALEX v.6.41 (Peakall and Smouse 2006) and compared between the flax core collection and the 30 Canadian cultivars across all identified stable QTL/markers. Significant differences between the allele frequencies were ascertained by the Kruskal-Wallis non-parametric test (Kruskal and Wallis 1952).

4.4 Results

4.4.1 Phenotypic data

All seed quality traits showed significant genotype (G), location (L) and year (Y) effects ($P < 0.001$; Appendix XI), although G explained a much larger percentage of the

phenotypic variation (33.3-90.6%) than L (1.2-26.5%) and Y (0.5-7.3%). Most of the genotype by environment (GE) interactions (G*L, G*Y, L*Y and G*L*Y) were significant and accounted for up to 10% of the seed quality traits variation. The location means, standard deviations, ranges, *H* and the correlations exhibited by the seed quality traits are summarized in Table 4.1. In MB, *H* ranged from 0.87 to 0.99, while in SK, it ranged from 0.73 to 0.98, with a lower mean (0.89) than MB (0.95), indicating that the repeatability between years was more consistent in MB than in SK. LIN and IOD were highly correlated at both locations (MB = 0.87, SK = 0.76; $P < 0.001$). Highly significant negative correlations were observed between the other FAs and IOD. Most of the correlations between FAs were significant and negative. OIL was positively correlated with PAL at both locations and with STE and OLE in SK but negatively correlated with LIO and IOD in SK.

4.4.2 Linkage disequilibrium

As shown in Appendix XII, syntenic r^2 (estimated LD for the loci on the same LG) was predominant on LGs 3, 8, 12 and 14, while LGs 1 and 10 showed r^2 close to background level. Blocks of LD among unlinked loci, which can produce false positive associations were also identified suggesting that the kinship matrix used in the MLM could be used to control false positive LDs (Yu et al. 2006).

Table 4.1 Mean ± standard deviation, range, broad sense heritability (*H*) and correlation coefficients of seven seed quality traits in the flax core collection evaluated in six environments.

Trait	Location	Mean ± SD	Min-Max	*H*	OIL	PAL	STE	OLE	LIO	LIN	IOD
OIL	MB	41.6±1.9	33.4-49.7	0.87	-						
	SK	43.3±2.3	32.8-52.3	0.87							
PAL	MB	5.7±0.7	3.3-9.2	0.96	0.21*	-					
	SK	5.4±0.6	3.3-8.4	0.90	0.39*						
STE	MB	4.7±1.2	2.3-11.9	0.97	0.0*n.s.*	0.35*	-				
	SK	4.0±0.9	2.2-9.1	0.95	0.24*	0.25*					
OLE	MB	23.8±3.7	15.3-43.9	0.93	0.03*n.s.*	0.07*n.s.*	0.34*	-			
	SK	18.1±2.9	11.7-35.9	0.90	0.22*	0.11*	0.38*				
LIO	MB	13.6±4.5	4.9-69.2	0.99	-0.06*n.s.*	-0.12*	-0.18*	-0.30*	-		
	SK	14.6±4.5	6.6-70.0	0.98	-0.20*	-0.02*n.s.*	-0.17*	-0.23*			
LIN	MB	52.2±5.3	3.6-65.4	0.96	-0.01*n.s.*	-0.19*	-0.36*	-0.54*	-0.58*	-	
	SK	57.9±5.0	4.7-68.0	0.96	-0.05*n.s.*	-0.18*	-0.29*	-0.46*	-0.73*		
IOD	MB	180.7±8.4	143.1-200.3	0.95	-0.03*n.s.*	-0.38*	-0.63*	-0.78*	-0.14*	0.87*	-
	SK	192.0±8.0	134.4-208.4	0.73	-0.13*	-0.31*	-0.50*	-0.58*	-0.33*	0.76*	

* Significant at $P < 0.001$; *n.s.* = non-significant

Oil content (OIL), palmitic acid (PAL), stearic acid (STE), oleic acid (OLE), linoleic acid (LIO), linolenic acid (LIN), and iodine value (IOD)

4.4.3 AM analysis

The average relative kinship between any two genotypes was 0.023, and 80% of the pairwise kinship comparisons ranged from 0 to 0.05 (Appendix XIII). As depicted by the cumulative P-P plots (Appendix XIV), numerous spurious associations for all traits were observed with the GLM (*Q*). This model was characterized by an excess of small *P*-values indicating spurious associations. On the other hand, the GLM (PCA) over corrected the majority of the small *P*-values with few higher *P*-values departing at the very end of the expected distribution. The MLMs (*K*) and (*Q* + *K*) performed similarly for the seven seed quality traits with their observed *P*-values deviating the most from the expected ones for OIL, PAL, STE, OLE, LIO and IOD, indicating that inclusion of the *Q* matrix brought little or no improvement to the AM model. Nevertheless, they displayed a better distribution of *P*-values for LIN (Appendix XIV). The MLM (PCA + *K*) had the smallest deviation from the expected distribution for all seed quality traits. The three first PCAs in combination with the *K* matrix were sufficient to control the majority of the potential false positive associations created by population and family structure. As a result, the P-values generated by the MLM PCA + K were retained for posterior analyses.

4.4.4 QTL contributing to seed quality traits

AM was conducted on OIL, PAL, STE, OLE, LIO, LIN and IOD across six environments in the Canadian Prairies. The genomic distribution and number of significant markers, candidate QTL and their phenotypic contribution to seed quality traits are summarized in Fig. 4.1, Tables 4.2 and 4.3 and Appendix XV.

Nine QTL were detected for five seed quality traits. The QTL with the largest effects were *QIod-LG8.1, QLin-LG5.2*, and *QOil-LG9.1* for IOD, LIN and OIL, respectively (Table 4.3). No QTL were detected for PAL and OLE but marker Lu2046 on LG2 and marker Lu2555 on LG6 explained 8.4 and 3.9% of the variation, respectively, with one allele contributing significantly to PAL and OLE as described in the next section (Fig. 4.2b, d). Several QTL and markers co-located within the same chromosomal regions such as those for LIO and LIN on LGs 3, 5 and 12 and LIO, LIN and IOD on LG8 (Fig. 4.1).

4.4.5 Allelic effects of stable associations

Some alleles were significantly associated with positive improvements of the traits. For example, the 270bp allele of Lu181 significantly increased OIL by an average of 1.3% ($P < 0.001$) across the six environments tested (Fig. 4.2a). For Lu2534, the 312bp allele had the largest effect on PAL increasing the value by ~1% over the average of the other alleles ($P < 0.001$; Fig. 4.2b). For STE, the 356, 358 and 360bp alleles of Lu146 had significantly larger effect than the other two alleles (Fig. 4.2c). An increase of 2.3% ($P < 0.001$) in OLE was associated with the 217bp allele of Lu2555 (Fig. 4.2d). Lu3262 explained ~8% of the variation for LIO with the 195bp allele increasing the trait by 0.9% (Fig 4.2e). The same allele was also associated with an increase in LIN of 1.3% (Fig. 4.2f). A significant positive effect of the 241bp allele of Lu2102 increased IOD by 9.5 units (Fig. 4.2g) ($P < 0.001$).

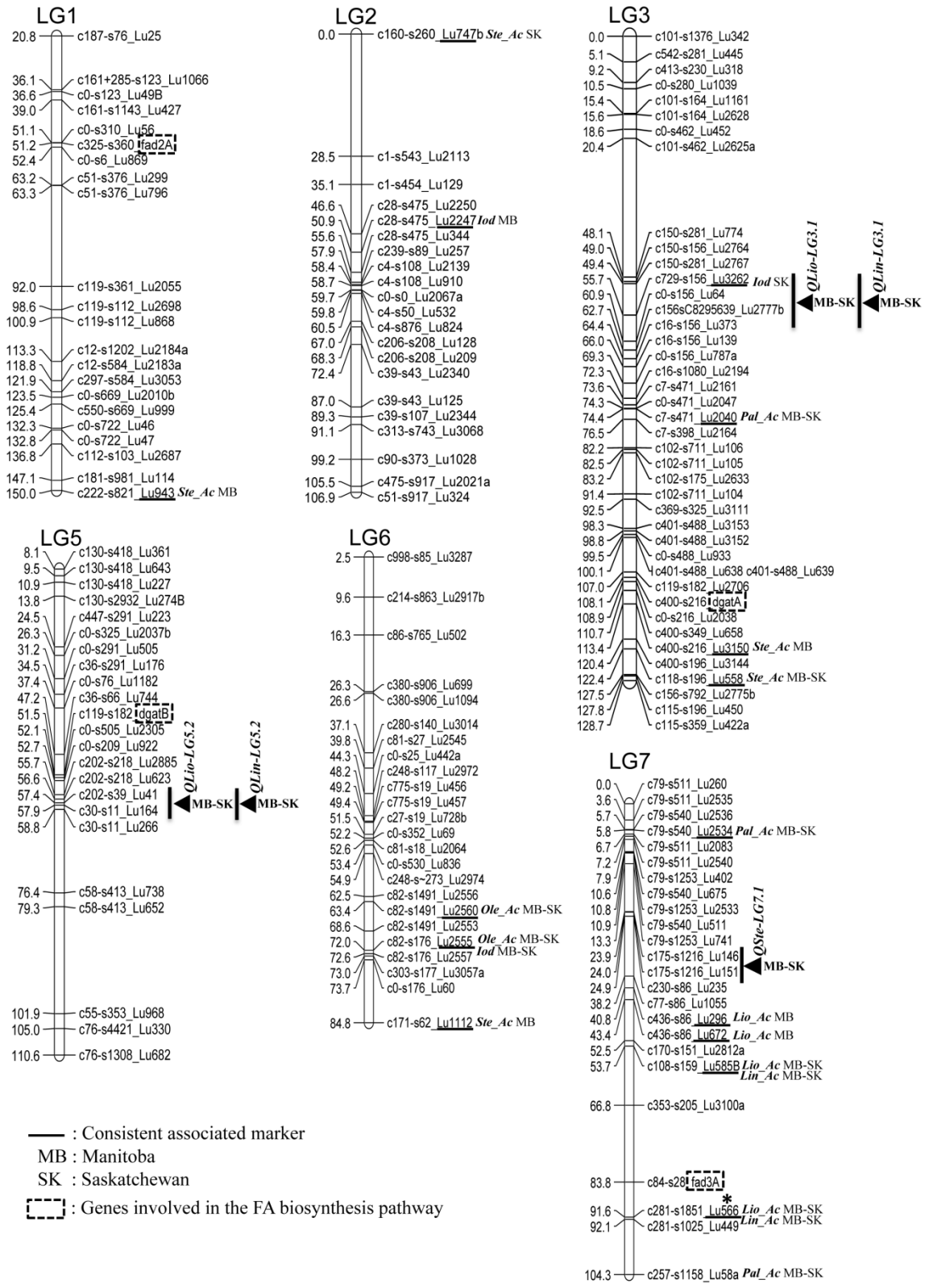


Fig. 4.1 Consensus genetic map of flax (Cloutier et al. 2012b) showing the location of the stable associated markers and candidate QTL for seven seed quality traits in linseed. Asterisks (*) indicate QTL previously reported (Cloutier et al. 2011).

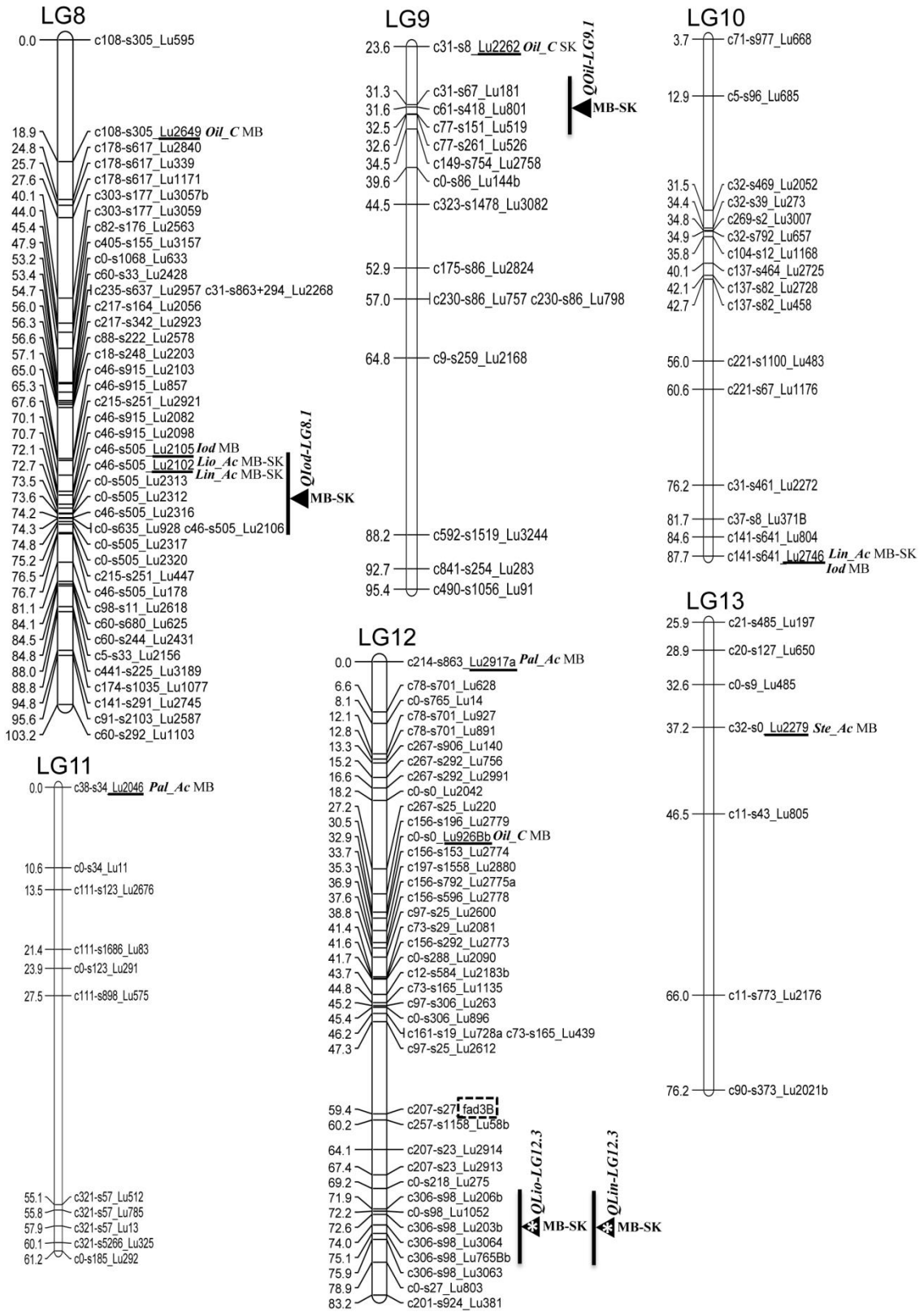


Fig. 4.1 Continued.

Table 4.2 Summary of significant markers and candidate QTL associated with seven seed quality traits in linseed identified using the MLM (PCA + *K*). QTL details can be found in Table 4.3 and Appendix XV.

	$-Log_{10}$ (P) threshold	No. of significant markers	% phenotypic variance (R^2)[a]	No. of candidate QTL	% phenotypic variance (R^2)[a]
Manitoba (MB)					
Oil content	3.3	7	16.8	1	3.7
Palmitic acid	3.0	4	11.4	0	0
Stearic acid	3.4	10	42.2	1	13.2
Oleic acid	3.6	2	5.5	0	0
Linoleic acid	3.6	15	40.6	3	34.3
Linolenic acid	3.6	12	29.5	3	25.6
Iodine value	3.6	6	12.1	1	5.6
Saskatchewan (SK)					
Oil content	3.5	3	13.8	1	12.8
Palmitic acid	3.1	3	5.3	0	0
Stearic acid	3.2	7	31.9	1	8.2
Oleic acid	3.8	2	6.4	0	0
Linoleic acid	3.5	13	38.1	3	31.8
Linolenic acid	3.5	12	30.2	3	27.0
Iodine value	3.1	5	13.3	1	5.8
Both locations					
Oil content	3.3	2	9.3	1	9.3
Palmitic acid	3.0	2	3.2	0	0
Stearic acid	3.2	3	11.7	1	19.6
Oleic acid	3.7	2	6.2	0	0
Linoleic acid	3.5	13	37.4	3	23.5
Linolenic acid	3.5	12	30.3	3	20.7
Iodine value	3.2	2	7.4	1	6.5

[a]Total phenotypic variation explained by the associated markers and candidate QTL

Table 4.3 Stable candidate QTL associated with seed quality traits identified at both Manitoba (MB) and Saskatchewan (SK) locations

Trait	Contig-Scaffold-Marker	Allele size (bp)	LG	Position	-Log_{10} (P)	QTL	Size (cM)	R^2 (%)	LD (r^2)[a]	Effect	IPCA1	ASV
OIL	c31-s67_Lu181	270	9	31.34	3.73	*QOil-LG9.1*	1.20	7.56	0.27	1.33**	-1.062	2.38
STE	c175-s1216_Lu146	354	7	23.95	6.23	*QSte-LG7.1*	0.01	19.68	0.71	1.67**	-0.241	0.40
LIO	c729-s156_Lu3262	217	3	55.74	5.10	*QLio-LG3.1*	8.70	6.60	0.24	1.09**	-0.701	1.67
	c30-s11_Lu164	211	5	57.89	3.52	*QLio-LG5.2*	0.90	3.31	0.11	0.43*	1.239	2.16
	c306-s98_Lu765Bb	null	12	75.12	8.40	*QLio-LG12.3*[b]	3.20	13.6	0.93	0.90*	0.489	0.78
LIN	c729-s156_Lu3262	217	3	55.74	5.57	*QLin-LG3.1*	8.70	5.33	0.24	1.24**	0.501	1.09
	c202-s39_Lu41	323	5	57.36	5.99	*QLin-LG5.2*	0.90	9.31	0.11	1.79**	0.302	1.04
	c306-s98_Lu765Bb	null	12	75.12	4.86	*QLin-LG12.3*[b]	3.20	6.06	0.93	0.63*	-0.890	1.16
IOD	c46-s505_Lu2102	241	8	72.74	4.23	*QIod-LG8.1*	1.60	9.35	0.22	9.31**	0.807	1.05

[a] Strength of the physical linkage between markers ranges from 0 (no linkage or no correlation between alleles at different loci) to 1 (total linkage or perfect correlation between alleles at different loci)

[b] Candidate QTL previously reported (Cloutier et al. 2011)

Significance of the allelic effects tested by Kruskal-Wallis non-parametric test *$P < 0.01$; **$P < 0.001$

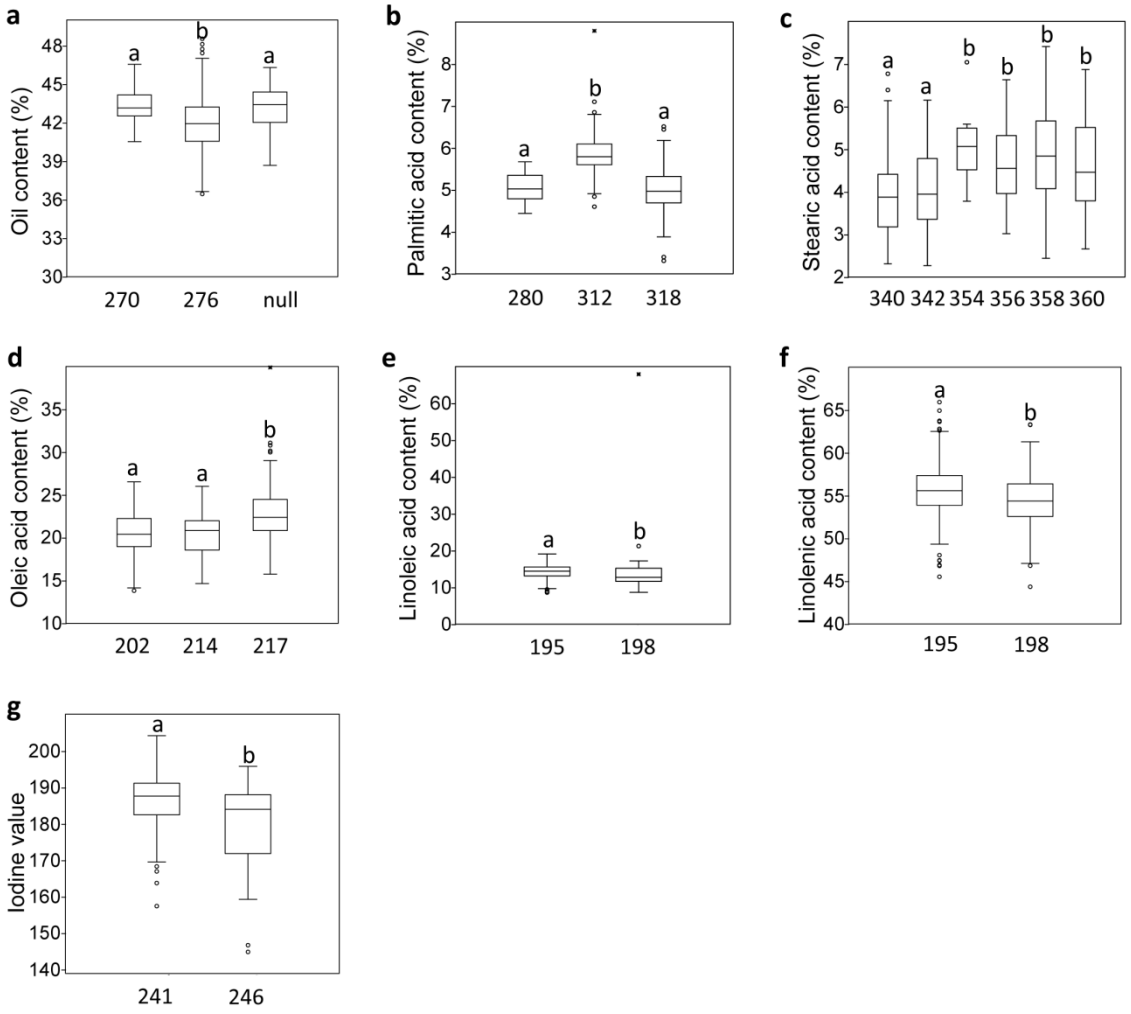


Fig. 4.2 Comparison of allelic effects of seven consistent associated markers with seed quality traits in linseed. **a** Lu181 associated with oil content **b** Lu2534 associated with palmitic acid content **c** Lu146 associated with stearic acid content **d** Lu2555 associated with oleic acid content **e** and **f** Lu3262 associated with linoleic and linolenic acid content **g** Lu2102 associated with iodine value. Bottom values represent the allele size in base pairs. Box plots followed by the same letter do not differ statistically according to the Kruskal-Wallis test ($\alpha = 0.01$).

4.4.6 QTL stability and QTL main effect

The AMMI analysis revealed that four of the nine candidate QTL identified for five seed quality traits were highly stable with IPCA1 values lower than ± 0.5 (Table 4.3). Also, all but three of the candidate QTL were stable or moderately stable with ASV in the range of 0.4 to 1.16.

The QQE biplot displays the average environment defined by the average PC1 and PC2 scores across environments (indicated by an open blue circle) (Fig. 4.3). The arrow passing through the biplot origin is called the AEC abscissa and points towards increasing QTL main effect. The AEC ordinate line, perpendicular to the abscissa and also passing through the biplot origin, indicates stability/instability. Highly unstable QTL have longer projections on the AEC abscissa irrespective of their direction. The LIN related candidate QTL/markers were highly stable because most of them landed on or very close to the AEC abscissa (Fig. 4.3). The intersection of the two axes defines the average QTL/marker main effect, and, as such, Lu203b, Lu2102, Lu206b, Lu566, *QLinLG12.3* and Lu585B had effects below average, while Lu2746, Lu2561a, *QLin-LG3.1*, *QLin-LG5.2*, Lu373 and Lu164 had the largest main effects on LIN across environments. In general, the QTL main effects showed by the QQE biplot were in agreement with the estimated phenotypic effect (Table 4.3, Appendix XV).

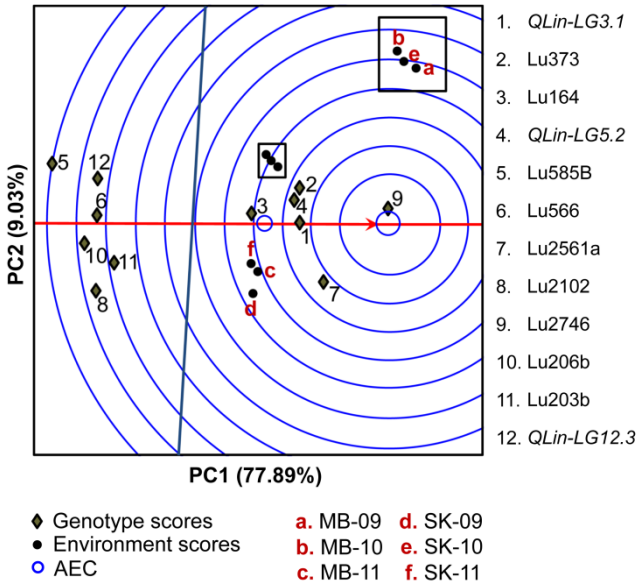


Fig. 4.3 QQE biplot for QTL main effect and QTL stability of linolenic acid content.

4.4.7 Frequency of QTL/marker alleles in the flax core collection and Canadian cultivars

Nine QTL/markers and 16 associated markers not part of a QTL but that explained at least 1% of the phenotypic variation were included in the analyses, totaling 25 QTL/markers, where some of them were associated with more than one trait (Table 4.3, Online Resource 6). 43 QTL/marker alleles were present in the 30 lines representing the Canadian cultivars (Appendix X) and 102 were present in the remaining 377 lines of the core collection, while the observed number of private QTL/marker alleles, which are alleles exclusively present in a group and absent in the other, was one and 77, respectively. After adjusting for sample size differences, the QTL/marker allelic richness was estimated at 43 and 71 in the Canadian cultivars and the core collection respectively, while the number of private QTL/marker alleles was 4 and 32, respectively. In the core collection, 65 of the observed QTL/marker alleles were rare (MAF < 0.05), whereas in the Canadian cultivars only 2 fell in this category (data not shown).

The frequencies of the favourable QTL/marker alleles (alleles with positive effects in increasing OIL and LIN) were not statistically different between the core collection and the Canadian cultivars for the seven seed quality traits (Kruskal-Wallis P = 0.437; Appendix XVI). Nevertheless, for most favourable QTL alleles, the Canadian cultivars had higher frequencies, indicating that Canadian flax breeders have been successful in pyramiding the best QTL alleles for seed quality traits. Five favourable alleles were absent in the Canadian cultivars but were also low in frequency in the core collection (Appendix XVI).

4.5 Discussion

Linseed oil and its FA profile define to a large extent its market end-use and value. Genetic progress can be accelerated once genetic diversity for the traits of interest and QTL architecture knowledge are available to breeders. In the present study, we described the application of AM using a core collection of 390 *L. usitatissimum* accessions for the identification of QTL underlying seed quality traits. This study establishes a framework to understand the quantitative nature of OIL and FA composition in linseed.

4.5.1 Phenotypic analysis

Significant GE interaction was observed for all seven seed quality traits, suggesting genotypic sensitivities to differences in environmental conditions (Appendix XI). In linseed, OIL and FA composition are affected by temperature during plant development (Casa et al. 1999; Fofana et al. 2006). Differences in planting dates and soil moisture can also affect OIL and FA composition in oil crops (van der Merwe et al. 2013). Fofana et al. (2006) showed that warmer and drier environmental conditions resulted in approximately 5% lower LIN compared to OLE and suggested that the fatty acid desaturase FAD2, which converts OLE into LIO, was more sensitive to environmental variations and therefore rate limiting. QTL for FA composition had already been linked to the FAD2 enzymes in flax (Cloutier et al. 2011). Our results are in line with this report where OLE was 5.7% higher in MB than in SK but LIN was higher in SK by the same percentage point. Historical meteorological data (30 year period) indicates that the MB location is warmer than the SK location, particularly during the growing season in 2010

and 2011 (Agriculture and Agri-Food Canada; http://climate.weather.gc.ca/advanceSearch/searchHistoricData_e.html).

Broad sense heritability (*H*) estimates were moderate to high with the phenotypic means and ranges reflecting the broad variation of the core collection and also indicating that a large proportion of the phenotypic variation was genetic. Genetic gain could be achieved through phenotypic selection; however, the correlations among seed quality traits exhibited complex relationships. The development of linseed cultivars with specific FA profiles could be better achieved through MAS for which a clear understanding of the genetic architecture of seed quality traits is needed.

4.5.2 AM analysis

The advantages of AM in identifying QTL for multiple traits in a single diverse population have been outlined (Gupta et al. 2005; Myles et al. 2009; Rafalski 2010). However, this approach sometimes suffers from an inflation of false positives due to population structure (Pritchard et al. 2000) and familial relatedness (Yu et al. 2006). Several linear and mixed models have been proposed to correct for the effect of both confounding factors (Pritchard et al. 2000; Price et al. 2006; Yu et al. 2006). In general, when population and family structures are present, the MLM is superior to the GLM (Myles et al. 2009) but in many cases, the best fitting model will depend on the dynamics of the association panel chosen. The *K* matrix can account for subtle population structure caused by familial relatedness, while the *Q* and PCA matrices control factors such as growth habit, market classes, geography, etc. PCA axes of variation have been shown to better adjust for allele frequency differences between subpopulations (Price et al. 2006;

Ma and Amos 2012). In our previous study, one of the two major STRUCTURE sub-groups clustered more than 90% of the fiber flax accessions, indicating that the inferred *Q* matrix mostly accounted for plant morphotype differences (Soto-Cerda et al. 2013) and hence, geographic differences present in the flax core collection might not be properly interpreted by the *Q* matrix we fit (Price et al. 2006). For all seven seed quality traits studied herein, the PCA + *K* model provided the best adherence to the expected cumulative distribution of *P* values (Appendix XIV), being superior to the *K* and *Q* + *K* models. This suggests that, in the case of linseed, the PCA matrix can better correct for population stratification, which turned out to also be computationally advantageous even with thousands of markers (Price et al. 2006).

4.5.3 Fatty acid QTL

Seed oils are composed primarily of triacylglycerols (TAGs), which are glycerol esters of FAs (Rao et al. 2008). The primary FAs in the TAGs of oilseed crops are 16-18 carbons in length and contain 0-3 double bonds where PAL, STE, OLE, LIO and LIN predominate (Rao et al. 2008). Only three FA-related QTL have been identified to date in flax: two co-located QTL, each associated with LIO, LIN and IOD and one affecting PAL (Cloutier et al. 2011). In the present study, we validated one of them, i.e., the co-located *QLio-LG12.3* and *QLin-LG12.3* (Fig. 4.1 and Table 4.3) located in the block of LD on LG12 (Appendix XII). Several markers and candidate QTL mapped close to genes involved in the FA biosynthesis pathway. Marker Lu3150 (LG3) associated with STE, mapped 5.3 cM from the acyl-CoA:diacylglycerol acyltransferase A (*dgatA*) gene (Fig. 4.1). Cloutier et al. (2011) mapped the gene using the microsatellite markers present in

the upstream region of the *dgat1* gene which was characterized from a bacterial artificial chromosome (BAC) clone. Highly significant associations between *DGTA1-2* and OLE and OIL have been reported in maize (Chai et al. 2012). A direct role for DGAT in STE is not obvious because DGAT-A and -B exert their main control in the final steps of oil assembly and are hypothesized to be a determining factor of OIL in higher plants (Weselake 2005). The associations with STE may be caused by the LD between the *dgatA* gene and the putative causative gene, a causal effect which could be resolved with a higher marker density. On the other hand, some of the oil assembly enzymes have been shown to have a preference for certain FAs (Sørensen et al. 2005). Such a selective mechanism could explain their indirect influence on the FA composition because most of the FAs will be assembled in TAGs.

Marker Lu566 (LG7) associated with LIO and LIN co-localized to the same region as the *fad3A* gene, overlapping with the previously published QTL *QLio.crc-LG7* and *QLin.crc-LG7* (Cloutier et al. 2011), thus being a major candidate gene for the control of LIN. Three *fad3* genes have been identified in the flax genome: *fad3a* and *fad3b* from cultivar Normandy (Vrinten et al. 2005) and more recently *fad3c* (Banik et al. 2011). FAD3A and FAD3B are major enzymes controlling LIN content in linseed (Vrinten et al. 2005); they were mapped in a bi-parental population (Cloutier et al. 2011) and recently integrated into the consensus map of flax (Cloutier et al. 2012b). In linseed, DGATA has an enhanced specificity for α-18:3-CoA (Sørensen et al. 2005; Rao et al. 2008), hence, higher LIN could translate to higher OIL in favourable environments such as SK where LIN was 5.7% higher and OIL was 1.7% higher than at the MB location (Table 4.1).

The genetic architecture of the traits provides some insights into the detection of more QTL for FA composition as compared to OIL. Variations in FA composition are mainly determined by a small number of major genes including fatty acid elongases and desaturases, while the number of genes potentially involved in OIL is expected to be greater and also more sensitive to environmental variations (Honsdorf et al. 2010). The marker density also likely played a role. The 460 SSRs represent less than one third of the 1,500 estimated minimum markers necessary to tag all QTL, indicating that potentially many QTL remained undetected. Likewise, the flax morphotype i.e. oilseed and fiber flax, could negatively impact on the number of significant associations. When alleles segregate across multiple subpopulations, MLMs are more powerful but when they segregate in only one or a subset of the subpopulations or, when different alleles are present in the subpopulations, MLMs will fail to detect the associations entirely (Zhao et al. 2011). We cannot discard the potential effect of the fiber morphotype on seed quality traits associations because it is likely that the favorable alleles associated with these traits do not segregate homogeneously across sub-groups, or they could even be totally absent in the fiber accessions which have not been selected for these traits, consequently under powering the AM results. AM analysis conducted separately for the fiber and oilseed accessions could provide further insights in this regard.

The phenotypic correlations between traits were consistently reflected in the identification of common markers and candidate QTL (Fig. 4.1) as reported in other QTL studies (Bachlava et al. 2009; Honsdorf et al. 2010; Cloutier et al. 2011; Hamdan et al. 2012; Li et al. 2012). For example, the stable QTL defined by markers Lu2102 and Lu928 on LG8 (Fig. 4.1), was not only associated with IOD but also with LIN which

were positively correlated. Another candidate QTL between markers Lu206b and Lu765Bb on LG12 (Fig. 4.1), associated with both LIO and LIN, overlapped with the previously reported QTL *QLio.crc-LG16* and *QLin.crc-LG16* having significant negative correlations (Cloutier et al. 2011). Negative correlation between LIO and LIN has been observed in *Brassica napus* (Honsdorf et al. 2010) and common QTL affecting several FAs have also been reported in soybean (Bachlava et al. 2009; Xie et al. 2012) and safflower (Hamdan et al. 2012).

4.5.4 Marker/QTL effects and QTL stability

To maximize the initial impact of MAS in crops with a lack of molecular tools, such as linseed, the associated markers should be closely linked to the QTL and the mapped QTL should ideally have large effect and high stability. For example, the two QTL associated with LIO and LIN reported by Cloutier et al. (2011) were located in a confidence interval of 11.6 cM. In our study, we narrowed down those QTL to 3.2 cM and showed their high stability and high LD (Table 4.3). Improvement in linkage tightness translates into recombination probability reduction, thus creating better markers for MAS. Nevertheless, because large effect and highly stable QTL will be first fixed in breeding programs, large effect and environment specific QTL should also be targeted by breeders. For example, *QOil-LG9.1*increased OIL by 1.3% but exhibited higher instability than the other QTL (Table 4.3). Although our statistical threshold for linked LD was 0.1 which could be considered weak for effective MAS, seven of the identified candidate QTL showed moderate to high LD in the range of 0.22 to 0.93.However, the phenotypic variation explained by the same QTL differed between studies. In Cloutier et al. (2011), the QTL

associated with LIO and LIN explained 20% each of the variation, higher than the 13.6 and 6.1% reported in the present study. Many AM studies in humans have reported low R^2 values, labeling the remaining unexplained variation as the missing heritability (Myles et al. 2009). In *Brassica napus*, 57 significant markers explained up to 18% of the phenotypic variation for OIL (Zou et al. 2010), while in maize, 26 loci explained up to 83% (Li et al. 2013). There are several reasons for this. First, insufficient marker coverage where the causal polymorphism is not in perfect LD with the genotyped markers affects the detection power of AM leaving unexplained a higher percentage of the variation (Myles et al. 2009). Second, rare alleles with large effects remained undetected because they were excluded for statistical reasons (Breseghello and Sorrells 2006; Rafalski 2010). Third, traits controlled by large number of genes/QTL, each with small individual effects may escape statistical detection (Manolio et al 2009). Fourth, variation resulting from epistatic interactions between genes might also go undiscovered because epistasis can only be investigated practically in a sequential scan of major common loci (Storey et al. 2005). Finally, epigenetic variations are emerging as a major cause of the missing heritability (Rakyan et al. 2011). Epigenome-wide association studies are likely going to shed some light on the specific epigenetic mechanisms at play in phenotypic variation (Rakyan et al. 2011), and most interestingly their environmental and trans-generational stabilities. Bi-parental mapping has the power to detect the effects of rare alleles (Gupta et al. 2005). As such, high R^2 values reported by Cloutier et al. (2011) using a bi-parental cross of high LIN with low LIN, providing an extreme range of FA profiles, likely correspond to the mutant parental line major fatty acid desaturase rare alleles of large effect while in AM, the smaller R^2 values could correspond to common

variants of small effects from the same locus. Allele frequency differences for the same underlying locus between bi-parental populations and AM panels affect the explained phenotypic variation (Stich et al. 2008). The maximum proportion of the variance explained by a marker is observed for allele frequencies of 0.5, as expected in bi-parental populations such as recombinant inbred lines or F_1-derived doubled haploids. For an AM panel, the allele frequencies are expected to be considerably different from 0.5 especially when multi-allelic markers such as SSRs are used (Stich et al. 2008). As a consequence, the proportion of the variance explained by a marker is notably lower despite the same underlying allelic effect (Stich et al. 2008). In our study, the majority of the associated markers and candidate QTL explained < 5% of the phenotypic variation. Nevertheless, some candidate QTL explained up to 19 % of the phenotypic variation, and major QTL for OIL (8 %), STE (19.6 %), LIO (6.6 %) and LIN (9.3 %) were stable, making them suitable for MAS (Table 4.3; Fig. 4.3).

4.5.5 Frequency of QTL/marker alleles in the flax core collection and Canadian cultivars

Several reports indicate that Canadian linseed cultivars have been developed from a narrow genetic base (Fu et al. 2002, 2003; Cloutier et al. 2009) which is an impediment to further breeding progress. In the present study, the flax core collection showed abundant QTL allelic diversity with approximately 8 times more unique (private) alleles than the Canadian cultivar subgroup. However, the majority of these novel QTL alleles were rare, limiting their exploitation in AM, hence requiring different strategies for their efficient utilization. Among these potential strategies, optimal bi-parental mapping populations

could be designed using the comprehensive phenotypic and genetic characterization of the flax core collection. In addition, the joint use of linkage mapping and association models through the design of multiparent advanced generation intercross (MAGIC) or nested association mapping (NAM) populations can overcome the population structure issue (Rafalski 2010). These populations are advantageous from the point of view of increasing the frequency of rare alleles and balancing the overall allele frequencies, although the strong kinship relationships could be an impediment. However, the high kinship relationships among genotypes could be mitigated by MLM and exploited through genomic selection, a strategy complementary to AM which uses genome-wide marker information to model phenotypic traits and obtain estimated breeding values (Meuwissen et al. 2001).

4.6 Final remarks

The current study represents the first AM analysis in linseed. We identified 9 consistent QTL across six environments for seed quality traits and several stable markers providing a basis for further AM and fine mapping efforts aiming to understand the genetic architecture of seed quality traits in linseed. Although this study was somewhat limited with respect to marker density, novel QTL were mapped and several previously reported were validated. To realize the full potential of AM and of the flax core collection, whole genome re-sequencing of the entire core collection is under way to saturate the genetic map with hundreds of thousands of single nucleotide polymorphism markers. Validation of candidate QTL in bi-parental populations will guide the development of marketable linseed cultivars using MAS.

GENOMIC REGIONS UNDERLYING AGRONOMIC TRAITS IN LINSEED (*LINUM USITATISSIMUM* L.) AS REVEALED BY ASSOCIATION MAPPING

Braulio J. Soto-Cerda[1,2] · Scott Duguid[3] · Helen Booker[4] · Gordon Rowland[4] · Axel Diederichsen[5] · Sylvie Cloutier[1,2]

[1]University of Manitoba, Department of Plant Science, 66 Dafoe Road, Winnipeg, MB, R3T 2N2, Canada

[2]Cereal Research Centre, Agriculture and Agri-Food Canada, 195 Dafoe Rd, Winnipeg, MB, R3T 2M9, Canada

[3]Morden Research Station, Agriculture and Agri-Food Canada, 101 Route 100, Unit 100 Morden, MB, R6M 1Y5, Canada

[4]University of Saskatchewan, Department of Plant Sciences, College of Agriculture and Bioresources, 51 Campus Drive, Saskatoon, SK, S7N 5A8, Canada

[5]Plant Gene Resources of Canada, Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK, S7N 0X2, Canada

The author Braulio J. Soto-Cerda carried out the analyses, interpretation of data and co-wrote the manuscript. The major supervisor Dr. Sylvie Cloutier designed the study, supervised the work, and co-wrote the manuscript. Dr. Scott Duguid conducted the field trials in Manitoba. Drs. Helen Booker and Gordon Rowland conducted the field trials in Saskatchewan. Dr. Axel Diederichsen developed the flax core collection.

The manuscript was published in **Journal of Integrative Plant Biology 2013**, **doi:10.1111/jipb.12118.**

5.0 GENOMIC REGIONS UNDERLYING AGRONOMIC TRAITS IN LINSEED (*LINUM USITATISSIMUM* L.) AS REVEALED BY ASSOCIATION MAPPING

5.1 Abstract

The extreme climate of the Canadian Prairies poses a major challenge to improve yield. Although it is possible to breed for yield per se, focusing on yield-related traits could be advantageous because of their simpler genetic architecture. The Canadian flax core collection of 390 accessions was genotyped with 464 simple sequence repeat markers, and phenotypic data for nine agronomic traits including yield, bolls per area , 1000 seed weight , seeds per boll , start of flowering , end of flowering , plant height , plant branching and lodging collected from up to eight environments was used for association mapping. Based on a mixed model (principal component analysis (PCA) + kinship matrix (*K*)), twelve significant marker-trait associations for six agronomic traits were identified. Most of the associations were stable across environments as revealed by multivariate analyses. Statistical simulation for five markers associated with 1000 seed weight indicated that the favorable alleles have additive effects. None of the modern cultivars carried the five favorable alleles and the maximum number of four observed in any accessions was mostly in breeding lines. Our results confirmed the complex genetic architecture of yield-related traits and the inherent difficulties associated with their identification while illustrating the potential for improvement through marker-assisted selection.

5.2 Introduction

Linseed (*Linum usitatissimum* L.) is important for the oil and nutraceutical industries (Green et al. 2008). Its oil, characterized by a high concentration of omega-3 alpha linolenic acid (~55%), is widely recognized for its health benefits (Simopoulos 2000). A unique feature of linseed resides in the prospect of also commercializing its stems because they produce good quality fibres that have many end-uses (Czemplik et al. 2011) including paper, technical fibre and biofuels (Diederichsen and Ulrich 2009; Cullis 2011). In 2011, the total world production of linseed reached ~1.6 million tonnes, with Canada (~23%), China (~21%) and The Russian Federation (~14%) being the main producers (FAOSTAT 2013). Although Canada is the world's largest linseed producer and exporter (FAOSTAT 2013), linseed remains a minor crop, in part because its yield has been stagnating over the last decade, averaging 1.2 T/Ha compared to other oilseeds such as canola (rapeseed) that now reach 1.9 T/Ha (Statistics Canada; http://www.statcan.gc.ca).

Conventional breeding methods have been the cornerstone for linseed genetic improvement releasing new cultivars with durable resistance to diseases, agronomic fitness and greater yield stability (Green et al. 2008). However, the narrow genetic base used for the development of Canadian linseed cultivars (Fu et al. 2002, 2003; Cloutier et al. 2009), the scarce availability of related species to incorporate new variation, the lack of hybrid production systems (Green et al. 2008) and the limited genomic tools for molecular breeding (Cloutier et al. 2011, 2012a) have hampered yield and quality improvements, limiting linseed competitiveness.

Yield is the most important and complex trait in crops that shows correlations with other traits (Li et al. 2011b). In linseed, yield and its components such as 1000 seed

weight (TSW), seeds per boll (SPB) and bolls per area (BPA), are quantitatively inherited and controlled by many genes affected by multiple interactions with other genes and the environment (Shi et al. 2009; Parry and Hawkesford 2012; Cadic et al. 2013). An understanding of the genetic basis of yield-related traits is of practical value to breeders because such information assists in the design of efficient breeding strategies. This approach, focused on yield-related traits, has been embraced in oilseeds such as *Brassica napus* (Shi et al. 2009), soybean (Panthee et al. 2007; Liu et al. 2011) and maize (Huang et al. 2010b; Peng et al. 2011) focusing on the improvement and inheritance of yield-related traits for achieving greater yield. Other important agronomic traits such as flowering time (FL), plant height (PH), plant branching (PB) and lodging resistance (LDG) may also indirectly affect yield through various physiological mechanisms (Huang et al. 2010b; Li et al. 2011b), allowing crop phenology and plant architecture to be adapted to regional growing conditions, thus avoiding yield and quality losses (Duguid 2009). The estimation of the positions of quantitative trait loci (QTL) with consistent effects across environments for yield and its components and other agronomic traits is of central importance for marker-assisted selection (MAS) and, ultimately, for enhancing linseed competitiveness.

In oilseed breeding, most of the QTL contributing to yield and other agronomic traits have been identified through classical linkage mapping (Panthee et al. 2007; Shi et al. 2009; Huang et al. 2010b; Liu et al. 2011; Peng et al. 2011). Despite the proven usefulness of this technique to identify QTL involved in complex traits, the limited genetic diversity and recombination events accumulated in bi-parental populations impede the simultaneous identification of favourable alleles available to breeding

programs and the precision of the location of QTL, thus weakening MAS applications (Würschum 2012). Often presented as an alternative approach, association mapping (AM) makes use of all recombination events that have occurred during the history of a germplasm collection representing a broader genetic diversity and, consequently, leading to a higher mapping resolution and the simultaneous survey of a larger number of alleles (Flint-Garcia et al. 2003; Würschum 2012). In the last decade, AM has been successfully applied to crops (reviewed in Gupta et al. 2005; Soto-Cerda and Cloutier 2012), showing that faster breeding progress can be achieved (Myles et al. 2009; Cadic et al. 2013; Huang et al. 2013).

In 2009, the Total Utilization Flax GENomics (TUFGEN; http://www.tufgen.ca) project was initiated in Canada, generating a wealth of genomics resources with one of the main goals being applications to flax breeding (Cloutier et al. 2009, 2011, 2012a, b; Ragupathy et al. 2011; Venglat et al. 2011; Kumar et al. 2012; Wang et al. 2012a). The comprehensive characterization of the Canadian flax world collection preserved by Plant Gene Resources Canada permitted the assembly of the Canadian flax core collection of 390 accessions representing the diversity from 76 countries (Diederichsen et al. 2013). This valuable genetic resource ensures a cost-effective access to the diversity harboured in the whole collection of ~3,500 accessions (Diederichsen et al. 2013). Further molecular characterization of the Canadian flax core collection revealed its abundant genetic diversity, weak population and family structure and quantified its relatively fast genome-wide linkage disequilibrium (LD) decay, all positive attributes for AM studies (Soto-Cerda et al. 2013). In the present study, we carried out AM for yield, TSW, SPB, BPA, start of flowering (FL5%), end of flowering (FL95%), PH, PB and LDG on the

Canadian flax core collection assessed in Western Canada over four years. The objective of this research was to identify QTL contributing to these agronomic traits that could be capitalized upon to assist in breeding superior linseed cultivars with improved yield and consequently market competitiveness.

5.3 Materials and methods

5.3.1 Plant material, genotyping and field trials

The Canadian flax core collection assessed in this study contains 381 accessions selected by Diederichsen et al. (2013) and nine accessions of relevance to recent Canadian flax breeding programs. The 390 accessions were genotyped with 464 simple sequence repeat (SSR) markers (Roose-Amsaleg et al. 2006; Cloutier et al. 2009, 2012a; Deng et al. 2010, 2011) distributed across the 15 linkage groups of flax (Cloutier et al. 2012b). All accessions were evaluated during four years (2009, 2010, 2011and 2012) at the Morden Research Station, Morden, Manitoba (MB) and at the Kernen Research Farm located near Saskatoon, Saskatchewan (SK), Canada. A type-2 modified augmented design (MAD) (Lin and Poushinsky 1985) was used for the field experiments from which phenotyping data was collected for nine agronomic traits. Main plots were arranged in grids of ten rows and ten columns. Each main plot was divided into five parallel subplots (2m x 2m with 20 cm row spacing) with a plot control (cv. CDC Bethune) located in the center. Additional subplot controls (cvs. Hanley and Macbeth) were assigned to five randomly selected main plots.

5.3.2 Phenotyping of agronomic traits

Yield and its components including TSW, SPB, BPA were obtained by harvesting two 0.5 meter sections of row from the central part of each subplot. The boll weight from each 0.5 m rows was measured to obtain the BPA. Four 25-boll subsamples were counted for each 0.5 m row which were weighted and threshed. The seeds from each subsample were counted and weighted to obtain the SPB and TSW. FL5% and FL95% were recorded as the number of days between sowing and when 5% and 95% of the flowers had opened, respectively. Plant height (PH in cm) was recorded at maturity using the average of 10 plants located in the center of the subplots. Plant branching (PB) was evaluated according to von Kulpa and Danert (1962) using a 1-6 scale which describes PB as the ratio of the total stem length without side branches to that with side branches as follows: 1 = 1/1, 2 = 1/2, 3 = 1/3, 4 = 1/4, 5 = 1/5, 6 = 1/6. PB ratings of five and six correspond to the typical fibre flax with long stems and bolls only in the upper part of the plants while ratings of three and four correspond to intermediate flax or linseed. Lodging (LDG) was scored using a 1-7 scale where 1 = upright, 3 = intermediate and 7 = lodged. The number of environments in which each agronomic trait was assessed differed between traits as indicated in Table 5.1.

5.3.3 Statistical analysis

Adjusted data was obtained for each trait as previously described based on the MAD (You et al. 2013). Normal distribution of the adjusted agronomic trait data was tested using the Shapiro-Wilk test (Shapiro and Wilk 1965) and normal probability plots. Traits with significant deviation from a normal distribution were log-transformed prior to AM

analysis including FL95% (SK12), PH (SK11) and PB (MB09, SK10, MB11). The adjusted phenotypic values were used to estimate the variance components using the GLM procedure in SAS 9.1 (SAS Institute 2004) as described in You et al. (2013). Broad sense heritability (*H*) across years within location was estimated to elucidate the location effect on each agronomic trait as follows: $H = \sigma^2_G / [\sigma^2_G + (\sigma^2_{GE} / e) + (\sigma^2_e / e\ r)]$ where σ^2_G, σ^2_{GE}, σ^2_e, e and r correspond to the genetic variance, the genetic by environment interaction variance, the residual variance, the number of environments and the replications per environment, respectively. Pearson's correlation coefficients were calculated to express the relationships between agronomic traits.

5.3.4 Population structure and linkage disequilibrium

Population structure and LD analyses for this core collection were previously reported (Soto-Cerda et al. 2013). Briefly, the flax core collection was assessed with 259 mapped neutral SSR loci which indicated that all accessions were organized into two major groups (G1 and G3) and one admixed group (G2) with a weak population structure (F_{ST} = 0.09). G1 included mostly accessions from South Asia, Western Europe and South America, while G3 included accessions from North America and Eastern Europe (Soto-Cerda et al. 2013). A relatively fast genome-wide LD decay of approximately1 cM (r^2 = 0.1) was estimated. To determine whether the nine agronomic traits differed between the two major groups as a consequence of the population structure, we applied the Kruskal-Wallis non-parametric test (Kruskal and Wallis 1952). For the significantly different traits ($P < 0.05$), a general lineal model (GLM) was fitted to estimate the amount of phenotypic variation explained by the population structure as estimated by the

membership coefficient (*Q)* matrix and the principal component analysis (PCA), considering traits as dependent variables and *Q* and PCAs as fixed.

5.3.5 Association mapping

The adjusted phenotypic values of the agronomic traits were used for AM. Five AM models were tested in TASSEL 2.1 (Bradbury et al. 2007) including two GLMs and three mixed linear models (MLMs). The first GLM incorporated the *Q* matrix as the fixed covariate while the second used PCA (Price et al. 2006). The first MLM incorporated the kinship matrix (*K*) (Yu et al. 2006) as a random effect only, while the second and third used in addition the *Q* matrix and PCA as fixed covariates, respectively. The *Q* matrix was estimated using 259 mapped neutral SSRs (Soto-Cerda et al. 2013). The PCA matrix calculated in TASSEL 2.1 retained the first three components. The *K* matrix was constructed on the basis of 448 SSRs using SPAGeDi (Hardy and Vekemans 2002). All negative values between individuals were set to zero (Yu et al. 2006). The best AM model was selected using cumulative probability-probability (P-P) plots. For the AM analysis, only minor allele frequencies (MAF) of more than 0.05 were retained (Breseghello and Sorrells 2006).

Association mapping analyses for the agronomic traits were carried out for each year and location independently. Correction for multiple testing was performed using the estimated false discovery (*q*FDR) values (Benjamini and Hochberg 1995). The *q* values were calculated with the QVALUE R package using the smoother method (Storey and Tibshirani 2003). Markers with *q*FDR less than 0.01 in at least half of the tested environments were considered significant. For markers significantly associated with a

trait, a GLM with all fixed-effect terms was used to estimate the amount of phenotypic variation explained by each marker (R^2). Allelic effects of the significant marker loci were calculated as the difference between the average phenotypic values of the homozygous alleles with MAF greater than 0.05. The significant differences between the allele means were estimated by the Kruskal-Wallis non-parametric test (Kruskal and Wallis 1952) and visualized as box plots.

5.3.6 Stability and effect of significant markers

Marker effects were calculated as the difference between the average values of the two most contrasting homozygous classes in each environment (defined as location-year), and significance between allele means was evaluated using the Kruskal-Wallis non-parametric test (Kruskal and Wallis 1952). Marker stability was estimated using the additive main effect and multiplicative interaction (AMMI) model (Zobel et al. 1988; Gauch 1992) in GenStat 14 (VSN International 2011). Markers with a first interaction principal component (IPCA1) near zero are more stable than those with positive or negative values. The AMMI's stability values (ASV) (Purchase 1997) were calculated using the following formula:

$\text{ASV} = \sqrt{\frac{\text{SSIPCA1}}{\text{SSIPCA2}}(\text{IPCA1})^2 + (\text{IPCA2})^2}$, where SSIPCA1 and SSIPCA2 are the sum of squares of the interactions of the first and second principal component analyses, respectively. We defined ASV values in the range of 0 to 1, as indicative of high stability across environments. In addition, the stability and effect of associated markers/QTL were graphically displayed using the QQE (QTL main effect and QTL by environment

interaction) approach where the first two IPCAs were plotted in a QQE biplot (Yan and Tinker 2005) using GenStat 14.

5.4 Results

5.4.1 Agronomic traits

All agronomic traits showed significant genotype (G), location (L) and year (Y) effects ($P < 0.001$; Appendix XVII). Most of the genotype by environment (GE) interactions (G x L, G x Y, L x Y and G x L x Y) were significant, except for yield where only L xY was significant. The overall means, ranges, *H* and coefficient of variations are summarized in Table 5.1. In MB, *H* ranged from 0.15 to 0.83, while in SK, it ranged from 0.37 to 0.78, indicating that the repeatability was highly variable among the agronomic traits at both locations. Among the 36 possible correlations, 25 were significant at $P < 0.01$ (Table 5.2). Yield and its components were positively correlated with one another but they were negatively correlated with the phenological traits FL5% and FL95%, the morphological traits PH and PB and the LDG agronomic trait.

Table 5.1 Number of environments, descriptive statistics and broad-sense heritability (*H*) for the nine agronomic traits assessed in the Canadian flax core collection**.**

Trait	Env.	Mean	Range	C.V (%)	*H* (MB)	*H* (SK)
Yield (K/ha)	6	1312.1	565.2 - 2468.8	36.2	0.59	0.59
Bolls per area (bolls/m^2)	6	4134.8	1653.6 - 6482.8	22.8	0.41	0.49
Thousand seed weight (g)	6	5.1	2.7 - 8.4	3.9	0.75	0.76
Seeds / boll	6	6.2	3.5 - 8.1	11.5	0.63	0.63
Flowering 5% (days)	7	45.1	40.0 - 61.9	3.3	0.83	0.47
Flowering 95% (days)	7	51.2	45.9 -71.4	3.3	0.80	0.49
Plant height (cm)	6	51.3	28 - 92.9	11.8	0.63	0.76
Plant branching	4	3.4	1.7 - 5.3	23.1	0.15	0.78
Lodging	8	1.3	1.0 - 3.3	19.1	0.20	0.37

Table 5.2 Pearson correlations amongst the nine agronomic traits in the Canadian flax core collection.

Trait	Yield	BPA	TSW	SPB	FL5%	FL95%	PH	PB	LDG
Yield	-								
BPA	0.528**	-							
TSW	0.173**	-0.285**	-						
SPB	0.541**	0.272**	-0.123*	-					
FL5%	-0.111*	0.029	-0.361**	-0.323**	-				
FL95%	-0.108*	0.036	-0.352**	-0.347**	0.964**	-			
PH	-0.140**	-0.046	-0.361**	0.026	0.506**	0.497**	-		
PB	-0.073	0.007	-0.265**	-0.049	0.429**	0.416**	0.633**	-	
LDG	-0.134**	-0.005	0.094	-0.354**	0.005	0.007	-0.261**	-0.238**	-

* and ** indicate significance at $P < 0.01$ and 0.001, respectively

5.4.2 Association between population structure and agronomic traits

Due to different population sizes (G1 = 153; G3 = 211) and unequal variances within the two major groups for the agronomic traits, the Kruskal-Wallis test was applied as suggested by Lin et al. (2008). Only PH showed significant differences ($P = 0.03$) with G1 accessions being three centimeters taller than G3 accessions (Appendix XVIII).

Of the 92 fibre flax accessions of the core collection, 48 (36% of G1) clustered within G1 while 23 (12.8% of G3) belonged to G3, suggesting that although the coefficient of population differentiation (F_{ST}) was weak (0.09), the fibre morphotype could be the main factor responsible for the population structure of the flax core collection. We investigated the pattern of population structure within G1 and G3 separately and showed that both major groups were organized in two subpopulations ($Q \geq 0.7$) and one admixed subpopulation ($Q < 0.7$) (Appendix XIX). Within G1, the two subpopulations largely corresponded to the oil and fibre morphotypes, with 91% of the fibre accessions initially clustering within this group (Appendix XIX). Within G3, however, the two subpopulation clusters reflected their geographic distribution with no clear sub-clustering of the 23 fibre accessions (Appendix XIX). Thus, flax morphotype and geographic distribution constituted the main factors responsible for the population structure patterns observed in the Canadian flax core collection, with the *Q* matrix and the first three PCAs explaining 11.3% and 39% of PH variation, respectively.

5.4.3 AM analysis in the core collection and sub-groups

As depicted by the cumulative P-P plots generated using the 390 accessions (Fig. 5.1), numerous spurious associations for all traits were observed with the GLM (*Q*).

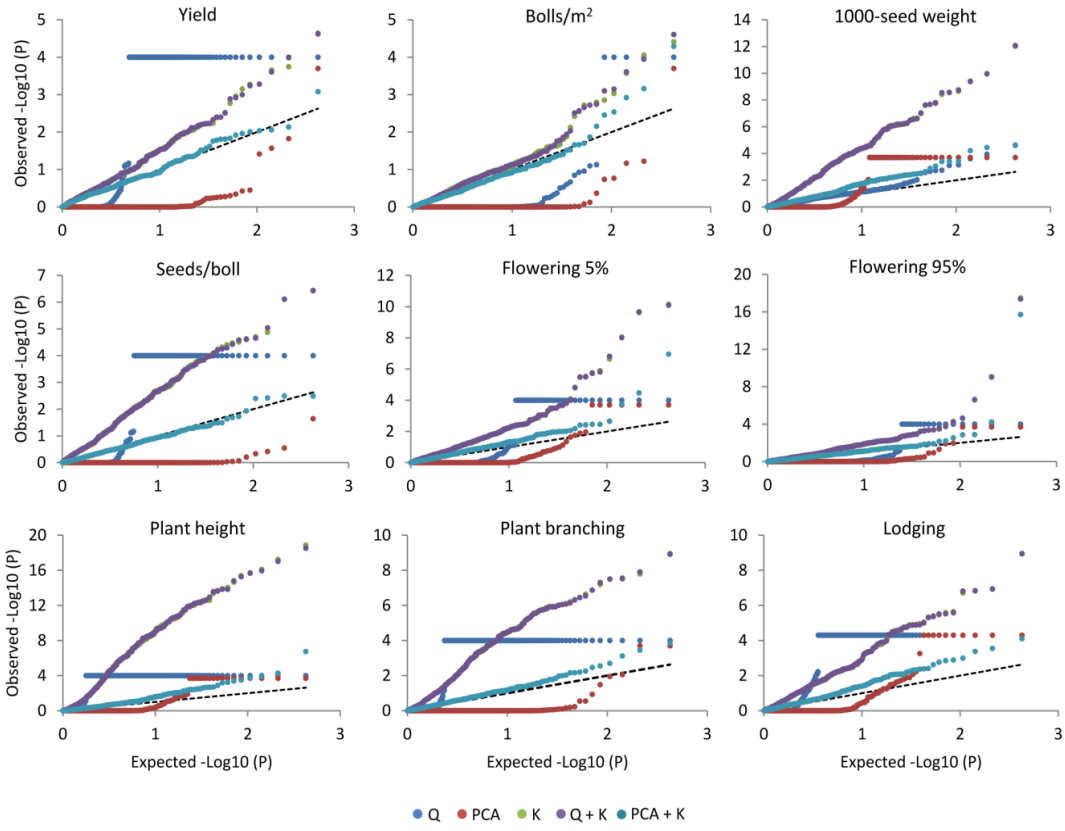


Fig. 5.1 Probability plots (P-P) of observed versus expected -Log_{10} (P) values for nine agronomic traits evaluated with five association mapping models. *Q* general linear model using the *Q* matrix, PCA general linear model using the principal component analysis matrix, *K* mixed linear model using the kinship matrix, *Q* + *K* mixed linear model using the *Q* and *K* matrices, PCA + *K* mixed linear model using the PCA and *K* matrices.

This model was characterized by an excess of small *P*-values causing spurious associations. On the other hand, the GLM (PCA) over corrected the majority of the small *P*-values with few higher *P*-values departing at the very end of the expected distribution.The MLMs *K* and *Q* + *K* performed similarly for the nine agronomic traits with their observed *P*-values deviating the most from the expected ones for TSW, SPB, PH, PB and LDG, indicating that inclusion of the *Q* matrix brought little or no improvement to the AM model. Nevertheless, they displayed a better distribution of *P*-values for BPA and FL95% (Fig. 5.1). The PCA + *K* MLM had the smallest deviation

from the expected distribution for all agronomic traits. The three first PCAs in combination with the *K* matrix were sufficient to control the majority of the potential false positive associations created by population and family structure. Therefore, the PCA + *K* model was selected to conduct AM for the nine agronomic traits in the core collection.

MLMs may overcompensate when traits are correlated with population structure, leading to false negatives (Zhao et al. 2011). Because up to 39% of the variation for PH was explained by population structure, we conducted AM for this trait within G1 and G3 separately. The P-P plot of G1 showed an improvement for the *K* and *Q* + *K* models, with the latter performing as well as the PCA + *K* (Appendix XIX). On the other hand, the P-P plot of G3 exhibited a better performance for the *Q* + *K* model only, the PCA + *K* being the most suitable. Thus, AM model comparisons indicated that conducting subpopulation-independent AM analyses partially alleviated the effect of population structure within G1 but did not correct it for G3, making it necessary to consider population structure as a fixed covariate. Hence, AM analyses for PH were conducted using the *Q* + *K* and PCA + *K* models.

5.4.4 Marker-trait associations

After removing alleles with MAF of less than 0.05, 37 SSR markers became monomorphic, leaving 427 polymorphic loci for the AM analyses. Using the PCA + *K* model, a total of 12 significant marker-trait associations (qFDR < 0.01) were identified as significant in at least half of the environments tested. They corresponded to 10 different markers distributed across six linkage groups (LGs). The majority of these associations

remained significant even after Bonferroni correction (0.05/427 = 1.17E-4) (Table 5.3). Numerous other significant associations were detected but they were not consistent in at least half of the environments. This was the case for yield, SPB and BPA, although six markers were associated with these traits in one or more of the environments. A total of five significant markers were associated with TSW, together explaining aproximately30% of the phenotypic variation for the trait (Table 5.3). Marker Lu943 was associated with FL5%, FL95% and PH, in agreement with their positive and significant correlations (Table 5.2). LG6 markers Lu2560 and Lu2564 located 0.7 cM apart formed a candidate QTL for LDG. For PH AM analyses, no additionalassociations were identified. However, for G1, marker Lu2067a associated with PB was correlated with PH (r = 0.633), and showed associations in two of the six environments evaluated.

5.4.5 Allelic effects of significant markers

Some of the alleles significantly improved TSW. For example, the 289bp allele of Lu526 significantly increased TSW by an average of 1.02 g (P = 8.5 E-13) across the six environments tested (Fig. 5.2a). For Lu2532, the 270bp allele had the largest effect, increasing TSW by 1.91 g (P = 1.7 E-6) over the 280bp allele and 1.3 g (P = 0.003) over the 282 bp allele (Fig. 5.2b). The 271bp allele of Lu943 significantly shortened FL5% by 2.13 days (P = 1.64 E-9) compared to the other two alleles (Fig. 5.2c). These allelic differences carried through to FL95% (Table 5.4). A reduction of up to 23.7 cm (P = 2.2 E-13) in PH was associated with the 241bp allele of Lu316 compared with the 223 bp allele (Fig. 5.2d). However, this large allelic effect can be inflated by the higher PH of the fibre accessions, where the 223bp allele was present in 33% of the fibre morphotype and

Table 5.3 Marker loci significantly associated with thousand seed weight (TSW), start of flowering (FL5%), end of flowering (95%), plant height (PH), plant branching (PB) and lodging (LDG), and their explained phenotypic variance (R^2).

Trait	Marker	LG (cM)[1]	MB09 (P value)	MB10 (P value)	MB11 (P value)	MB12 (P value)	SK09 (P value)	SK10 (P value)	SK11 (P value)	SK12 (P value)	R^2 (%)
TSW	Lu2164	3 (76.5)	N.E	*n.s*	*n.s*	1.61 E-4	N.E	**7.50 E-5**	**1.10 E-8**	**1.10 E-4**	0.50
	Lu2555	6 (72.0)	N.E	*n.s*	*n.s*	1.78 E-4	N.E	7.10 E-4	1.24 E-4	6.51 E-4	0.72
	Lu2532	7 (2.7)	N.E	*n.s*	*n.s*	**1.53 E-5**	N.E	**9.60 E-5**	**2.36 E-6**	**7.90 E-5**	8.0
	Lu58a	7 (104.3)	N.E	*n.s*	*n.s*	3.92 E-4	N.E	*n.s*	**2.38 E-6**	1.90 E-4	5.5
	Lu526	9 (32.6)	N.E	**4.20 E-5**	*n.s*	**6.81 E-6**	N.E	2.27 E-4	**1.10 E-4**	*n.s*	15.2
FL5%	Lu943	1 (149.9)	*n.s*	**4.42 E-7**	**7.88 E-5**	*n.s*	N.E	*n.s*	**4.34 E-5**	**7.35 E-7**	7.1
FL95%	Lu943	1 (149.9)	*n.s*	**2.60 E-5**	**8.94 E-5**	*n.s*	N.E	*n.s*	**8.74 E-5**	**4.90 E-6**	7.6
PH	Lu943	1 (149.9)	N.E	N.E	1.31 E-4	*n.s*	**1.01 E-4**	*n.s*	*n.s*	2.31 E-4	4.6
	Lu316	unknown	N.E	N.E	**1.15 E-5**	**9.23 E-5**	*n.s*	*n.s*	*n.s*	**1.62 E-5**	18.5
PB	Lu2067a	2 (59.7)	*n.s*	N.E	*n.s*	N.E	N.E	**9.08 E-5**	**3.35 E-5**	N.E	12.9
LDG	Lu2560	6 (63.4)	*n.s*	4.95 E-4	*n.s*	N.V	N.V	**5.73 E-5**	**1.38 E-18**	*n.s*	8.9
	Lu2564	6 (64.1)	1.53 E-4	8.74 E-4	**9.05 E-11**	N.V	N.V	*n.s*	1.20 E-4	*n.s*	7.1

[1] Linkage group and, in bracket, loci position in centiMorgan according to Cloutier et al. (2012b)

N.E trait not evaluated, N.V trait not phenotypically variable

Values in bold script are significant at *q*FDR < 0.01and after Bonferroni correction (0.05 / 427 = 1.17 E-4); those in normal script are significant at *q*FDR < 0.01; *n.s* non-significant

only 6% of the linseed morphotype while the 241bp allele was present in 31% of the linseed morphotype but only 7% of the fibre morphotype. The 205 bp allele of marker Lu2067a, increased PB up to 0.76 units compared with the 211bp allele (P = 2.03 E-8) (Fig. 5.2e). The null allele of Lu2560 decreased LDG by 0.34 units (P = 3.14 E-6) (Fig. 5.2f).

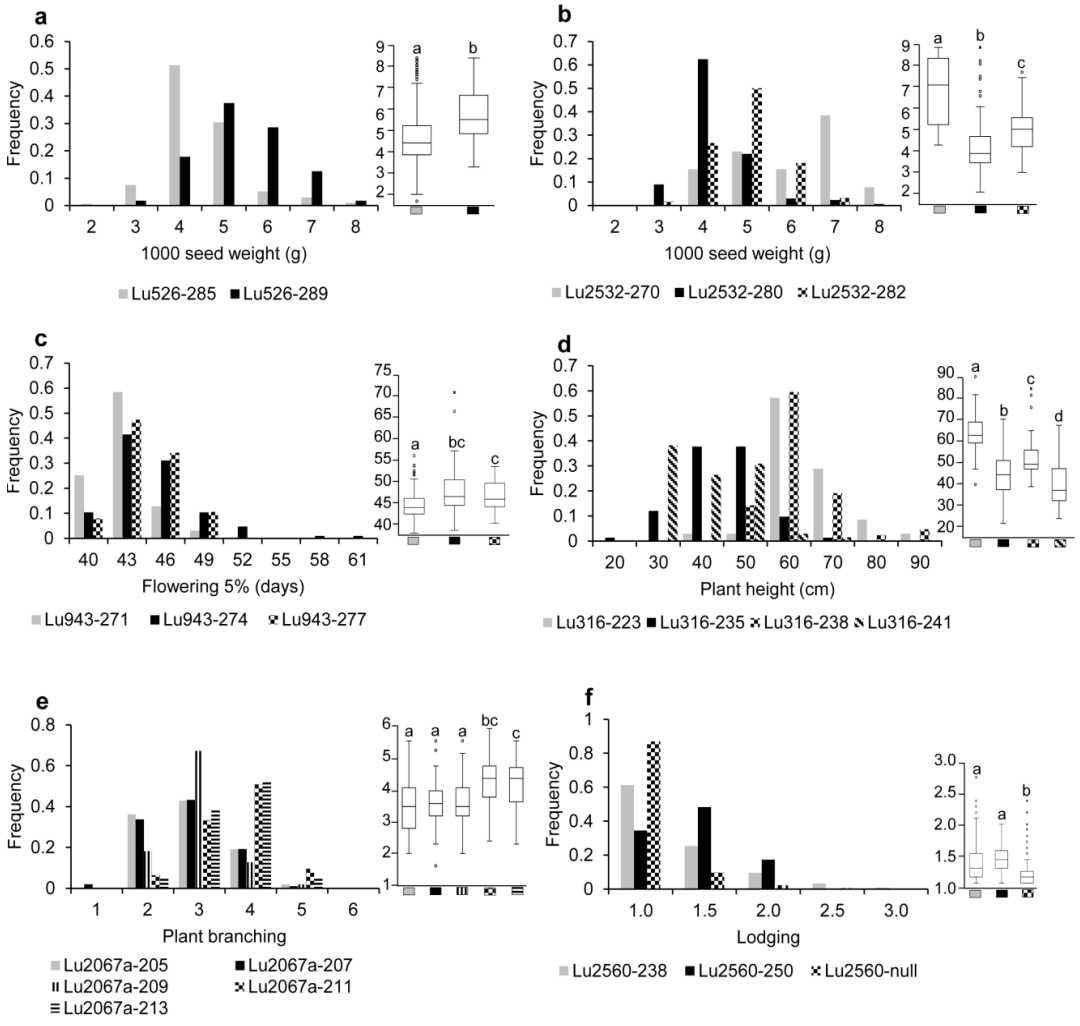


Fig. 5.2 Comparisons of allelic effects of six associated markers with agronomic traits in linseed. **a** Lu526 and **b** Lu2532 associated with thousand seed weight **c** Lu943 associated with start of flowering **d** Lu316 associated with plant height **e** Lu2067a associated with plant branching **f** Lu2560 associated with lodging. Box plots followed by the same letter do not differ statistically according to the Kruskal-Wallis test (α = 0.01).

5.4.6 Marker effect and stability

The AMMI analysis established that ⅓ of the marker trait-associations were highly stable with IPCA1 values close to ± 0.2 and that another third were moderately stable with values ranging from ± 0.25 to ± 0.6 (Table 5.4). The ASV stability parameter indicated that six marker trait-associations were highly stable with values ranging from 0.18 to 1.17. The QQE biplot displays the average environment defined by the average IPCA1 and IPCA2 scores across environments (indicated by an open circle) (Fig. 5.3a). The arrow passing through the biplot origin is called the AEC abscissa and points towards increasing marker/QTL main effect. The AEC ordinate line, perpendicular to the abscissa, indicates stability/instability. Highly unstable markers have longer projections on the AEC abscissa irrespective of their direction.

Table 5.4 Favorable alleles at the ten SSR loci associated with agronomic traits, their frequencies, phenotypic effects and stability.

Trait	Marker	Favorable allele (bp)	Freq. (%)	Effect	K-W test[a]	IPCA1[b]	ASV[c]
TSW	Lu2164	377	44.9	0.68 g[d]	1.9 E-3*	0.907	3.222
	Lu2555	202	47.9	0.85 g	2.1 E-12*	-0.411	1.446
	Lu2532	270	8.0	1.91 g	5.6 E-7*	-0.729	1.537
	Lu58a	209	72.5	0.72 g	3.1 E-3*	0.209	1.441
	Lu526	289	15.8	1.02 g	8.4 E-13*	0.023	1.178
FL5%	Lu943	271	60.8	-1.56 d	5.5 E-5*	-0.215	0.215
FL95%	Lu943	271	60.8	-2.15 d	1.2 E-9*	-0.181	0.181
PH	Lu943	271	60.8	-9.25 cm	8.4 E-9*	2.532	2.532
	Lu316	241	17.3	-23.7 cm	1.6 E-14*	-2.532	2.532
PB	Lu2067a	205	27.6	-0.76 u	1.5 E-9*	0.265	0.321
LDG	Lu2560	null	47.5	-0.34 u	4.7 E-8*	-0.557	0.558
	Lu2564	257	11.7	-0.28 u	6.4 E-4*	0.557	0.558

[a] *P* value for Kruskal-Wallis test for the allelic effect between favored alleles and others * $P < 0.01$
[b] First interaction principal component
[c] AMMI's stability values
[d] g = grams; d = days; cm = centimeters; u = units of PB and LDG scales, respectively

The markers associated with TSW varied in stability. For example, Lu2532 and Lu526 were more stable than Lu2555, Lu2164 and Lu58a (Fig. 5.3a). The intersection of the two axes defines the average marker/QTL main effect, hence, the latter three markers had effects below average; whereas, Lu2532 and Lu526 had the largest main effects on TSW across the six environments in which this TSW was tested (Table 5.4).

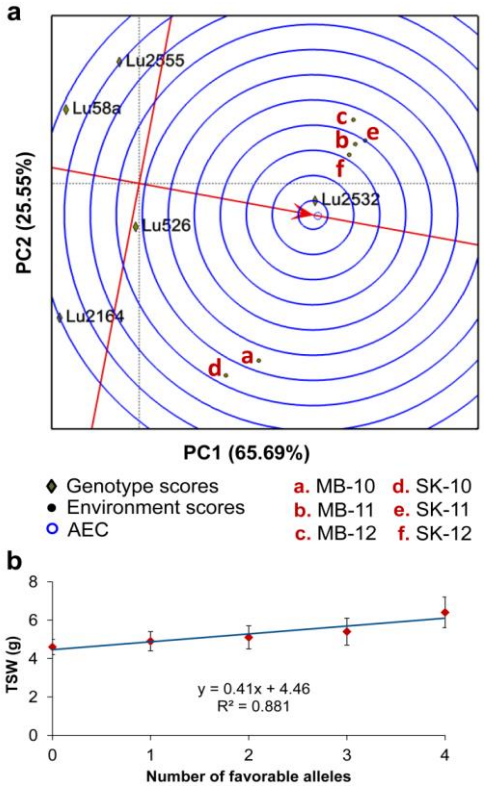


Fig. 5.3 Marker effect and stability of thousand seed weight. **a** QTL main effect and QTL-by-environtent interaction (QQE) biplot for marker/QTL main effect and marker/QTL stability of thousand seed weight **b** Linear regression analysis of thousand seed weight based on six environments.

Taking into consideration that approximately 300 accessions of the core collection are the linseed type, the favorable alleles of Lu2532 and Lu526, present in 31 and 62 accessions respectively, clearly demonstrate that they have not been the target of intensive selection by linseed breeders to date.

Linear regression analysis between TSW and the number of favorable alleles of associated markers showed a linear correlation, suggesting additive effects (Fig. 5.3b). No accession had all five favorable alleles but 10 accessions had four of them. Among these, only one U.S. modern cultivar (Maritime, mean TSW = 7.3 g) showed four alleles while the remaining nine were breeding lines including three belonging to the convar. *mediterraneum* characterized by its large seeds and high TSW (Fig. 5.4). The high yielding and broadly adapted Canadian cultivar CDC Bethune (mean TSW = 5.2 g) possesses only two of the five TSW favorable alleles.

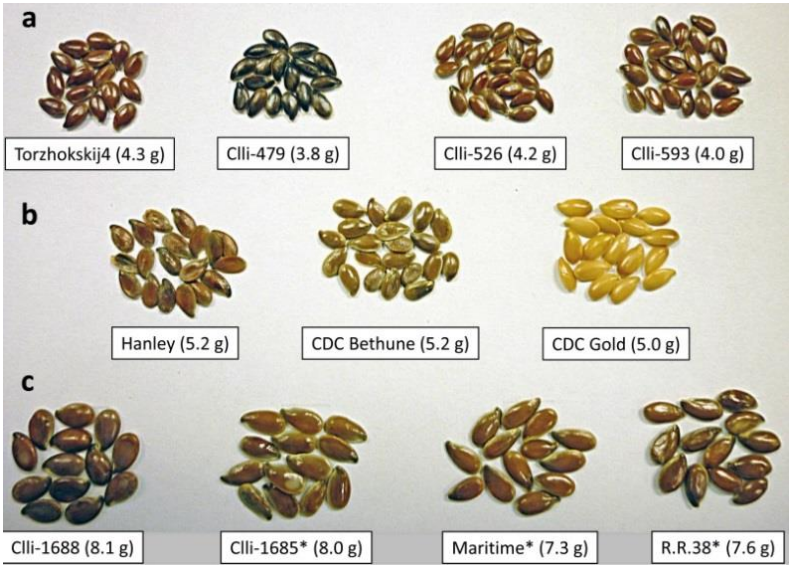


Fig. 5.4 Linseed accessions with different number of favourable alleles associated with thousand seed weight. **a** accessions with zero favorable alleles **b** Canadian cultivars with two favorable alleles **c** accessions with four favorable alleles. Values in brackets are the thousand seed weights for each accession and * indicates the accessions that belong to the convar. *mediterraneum*.

5.5 Discussion

Yield is a complex trait that can be broken down into its components which are in turn affected by other traits involving diverse pathways (Shi et al. 2009). For example, seed number, seed weight, flowering time, plant height and plant branching have all been

identified as affecting yield in rapeseed (Ishimaru 2003; Salamini 2003; Ashikari et al. 2005; Clark et al. 2006; Cockram et al. 2007). Phenotypic correlations and QTL analyses suggest that yield-associated traits tend to be clustered in the genome and have pleiotropic effects (Shi et al. 2009; Li et al. 2011a; Liu et al. 2011). Hence, understanding the genetic bases and relationships of yield-associated traits and agronomic traits in linseed through AM can provide the scientific background needed to devise breeding strategies that would permit and/or accelerate yield improvements beyond the 1.2 T/Ha achieved to date.

5.5.1 Agronomic traits

The analysis of variance showed that the genotype effect was highly significant for all nine traits, indicating that abundant and likely unexploited genetic diversity is harboured within the Canadian flax core collection (Table 5.1; Appendix XVIII). Yield, BPA and TSW had ranges that spanned five, four and three orders of magnitude, respectively (Table 5.1). GE interactions also contributed significantly to trait variations highlighting the need to identify stable germplasm across environments having favorable alleles (Zhang et al. 2010).

Broad sense heritability (*H*) is a suitable indicator of the trait repeatability and the proportion of trait variation accounted for by genetic factors. *H* varied largely between traits and locations. For example, the MB and SK locations had opposite effects on FL5%, FL95%, PB and LDG while their effects on yield-related traits followed similar trends (Table 5.1). Historical meteorological data indicates that the MB location is warmer and wetter than the SK location, this was particularly true during the growing

season of 2010 and 2011 ((Agriculture and Agri-Food Canada; http://climate.weather.gc.ca/advanceSearch/searchHistoricData_e.html). This complicates phenotypic selection of suitable parents with broad adaptation, the design of efficient breeding schemes and, ultimately, yield improvement.

Correlations among phenotypic traits are commonly observed in crops. Plant breeders need to consider trait correlations for the simultaneous improvement of numerous correlated traits or for reducing undesirable effects when the goal is to apply changes to one or a subset of the correlated traits (Chen and Lubberstedt 2010). Yield was positively correlated with its yield components and negatively correlated with FL5%, FL95%, PH and LDG (Table 5.2) suggesting that further yield improvement could come from the breeding of an early flowering, shorter linseed plant producing larger seeds per boll and more bolls per area. Similar phenotypic correlations among yield-related traits and other agronomic traits have been reported in soybean (Panthee et al. 2007), rapeseed (Honsdorf et al. 2010) and maize (Peng et al. 2011).

5.5.2 Association between population structure and agronomic traits

Correlations between population structure and variation for phenotypic traits have been reported (Camus-Kulandaivelu et al. 2006; Caniato et al. 2011; Zhao et al. 2011). In maize, a null allele of the *Dwarf8* (*D8idp*) gene associated with flowering time was found in high frequency among Northern Flint accessions but was rare in tropical accessions (Camus-Kulandaivelu et al. 2006). In sorghum, aluminium tolerance conferred by the *Sorghum bicolor* multidrug and toxic compound extrusion (*SbMATE*) gene was almost exclusive to West African genotypes (Caniato et al. 2011). Likewise in rice, several

height genes such as *Oryza sativa* BRI1-associated receptor kinase 1 (*OsBAK1*) and dwarf and gladius leaf 1 (*DGL1*) were population-specific and were only detected when no correction for population stratification was applied (Zhao et al. 2011). In our study, PH variation appeared to correlate with population structure caused by differences in plant morphotype because fibre flax and linseed differ considerably in morphology, anatomy, physiology and agronomic performance (Diederichsen and Ulrich 2009). Although incorporation of population structure covariate is important to control false positives in AM, a substantial fraction of the PH variation likely remained undetected as a consequence of the morphotypes in flax (Caniato et al. 2011).

5.5.3 AM analysis in the core collection and sub-groups

AM has demonstrated its power to detect QTL across multiple plant species and germplasm collections (reviewed in Gupta et al. 2005; Soto-Cerda and Cloutier 2012). However, a potential problem of AM resides in its inherent population stratification which is recognized as a source of spurious associations because phenotypic and genotypic variations end up highly correlated between subpopulations (Würschum 2012). To circumvent this limitation, a number of approaches have been suggested (Pritchard et al. 2000; Price et al. 2006; Yu et al. 2006). For all nine agronomic traits studied herein, the PCA + *K* model provided the best approximation to the expected cumulative distribution of *P* values (Fig. 5.1), being superior to the *K* and *Q* + *K* models. This suggests that, in the case of linseed, the PCA matrix can better correct for population stratification, in line with the larger PH variation explained by the first three PCAs, which

turned out to also be computationally advantageous even with thousands of markers (Price et al. 2006).

When alleles segregate across multiple subpopulations, MLMs are more powerful but when they segregate in only one or a subset of the subpopulations or, when different alleles are present in the subpopulations, MLMs will fail to detect the associations entirely (Zhao et al. 2011). Although we conducted AM for PH within each major group to minimize the confounding effects of flax morphotype and geographic distribution, it was necessary to use MLMs with population structure as covariate, but no significant associations were identified within the major groups. Because the simultaneous use of PCA and *K* matrices may result in overcorrection (Würschum 2012), additional PH QTL could be detected using biparental mapping populations developed from parents belonging to different subpopulations (Zhao et al. 2011) or, as recently proposed, through the design of multi-parent advanced generation intercross (MAGIC) or nested association mapping (NAM) populations (Mackay and Powell 2007; Yu et al. 2008).

5.5.4 Marker-trait associations

The number of significant associations varied considerably between traits, with no associations detected for yield per se, BPA and SPB, clearly emphasising the genetic complexity and high GE interaction of yield and its components (Shi et al. 2009). For example, five markers showed consistent associations with TSW, but more than 30 significant markers (qFDR < 0.01) were identified in at least one environment. These environment specific associations were detected for all traits. These associations may also result from weak LD between associated markers and QTL caused by (i) an insufficient

number of markers to cover all LD blocks across the genome (Würschum 2012), (ii) low trait heritability (Pasam et al. 2012) and (iii) the removal of rare alleles with large effects excluded from the analyses for statistical reasons (Breseghello and Sorrells 2006). In our study, marker density was likely a limitation considering that our LD analysis indicated that at least 1,500 markers would be required to provide the comprehensive coverage of the genome necessary for AM in the flax core collection (Soto-Cerda et al. 2013). Trait heritability likely negatively impacted marker-trait association detection because the observed *H* was low to moderate for the majority of the traits which also expressed significant location effect (Table 5.1, Appendix XVII). Other pitfalls include genomic regions close to fixation or totally monomorphic and that do not occur by chance, especially in large and diverse germplasm collections. We hypothesized that some of the 37 SSRs that became monomorphic after removal of the alleles with MAF of less than 0.05 have been selected during domestication or modern flax breeding such as the dehiscence trait, considering that they are shared across different populations (Kovach et al. 2007). As a result, they are totally uninformative using AM because the strength of LD mapping relies on polymorphisms between loci to estimate correlations between traits and their allele variants; thus, many potentially large-effect QTL were missed (Zhao et al. 2011). Genetic studies involving wild relatives, landraces and modern cultivars should help in elucidating this question (Vigouroux et al. 2002; Würschum 2012).

Yield improvement through yield components and related traits such as flowering time and plant morphology could be advantageous because of their simpler genetic architecture and higher stability than yield per se (Peng et al. 2011). In rapeseed, 785 QTL for eight yield-related traits were identified across 10 environments, but only 85

QTL for yield, of which none was consistent across environments (Shi et al. 2009). Exploiting the phenotypic correlations between yield-related traits can facilitate the pyramiding of favourable alleles because correlations may indicate linkage or pleiotropy (Li et al. 2011a; Zhao et al. 2011; Zhang et al. 2012). PH is an important developmental and yield-related trait and many genes regulating PH have been shown to affect harvest index and yield in rice (Xue et al. 2008; Xing and Zhang 2010), and yield and flowering time in soybean (Liu et al. 2011). The seemingly pleiotropic effect of the 271bp allele of Lu943 on FL5%, FL95% and PH illustrates the feasibility of developing short early flowering linseed cultivars with apparently no yield penalties using pleiotropic QTL (Li et al. 2011a). Similarly, TSW is an important yield component determining yield in crops (Li et al. 2011a; Liu et al. 2011; Wang et al. 2012b); thus, the combined selection of the five favorable alleles associated with TSW is a readily applicable strategy involving indirect yield improvement through yield components (Shi et al. 2009; Wang et al. 2012b).

5.5.5 Marker effect and stability

The majority of the associated QTL detected in biparental populations explained larger proportions of the variance than those detected in AM studies (Stich et al. 2008; Honsdorf et al. 2010; Pasam et al. 2012). Conversely, bias of biparental populations leads to an overestimation of the QTL effect, especially in small populations (Melchinger et al. 2004). In our study, the variance explained by the associated markers ranged from 0.5-18.5% (Table 5.3). Although no comparisons can be made with the non-existing previous QTL studies in flax for agronomic traits, these estimates are likely minimum estimates of

the real QTL effects because incomplete LD between marker and QTL leads to an underestimation of the variance explained by the QTL (Honsdorf et al. 2010; Würschum 2012). Comparable results between biparental mapping population QTL analysis and AM should be observed when LD is perfect ($r^2 = 1$) and the same alleles segregate in both populations (Myles et al. 2009). Even if LD was perfect, underestimation of the phenotypic variance could ensue from allelic frequency differential in the AM population (Stich et al. 2008). The maximum proportion of the variance explained by a marker is observed for allele frequencies of 0.5, as expected in biparental populations such as recombinant inbred lines or F_1-derived doubled haploids. For a germplasm collection, the allele frequencies are expected to be considerably different from 0.5 especially when multi-allelic markers such as SSRs are used (Stich et al. 2008). Thus, the proportion of the variance explained by a marker is notably lower despite the same underlying allelic effect (Stich et al. 2008). As a result, when AM is conducted with suitable marker density and the phenotypes are measured in representative environments, the variance explained by the associated markers should provide a more accurate estimation of the impact that the favorable alleles will have in a breeding program.

Quantitative trait loci with major effects and stable expression across environments and genetic backgrounds are better for MAS. Associations were declared only for markers significant in at least half of the tested environments and, using multivariate analyses, we estimated their stability and effects (Table 5.4, Fig. 5.3a). This approach enabled the identification of MAS candidate markers such as Lu2532, Lu526 and Lu943 that exhibited both high stability and large effects on TSW and flowering

traits. Other associated markers identified herein may also be useful for breeding because they all had moderate stability, although few had marginal R^2 values.

Molecular breeding aims to select the most valuable genotypes or alleles and to combine them in developing a desirable cultivar (Zhang et al. 2012). The identification of favorable alleles helps in selecting parents for crosses to ensure the pyramiding of the maximum number of favorable alleles in the best genetic background. In rice, linear correlation between TSW and favorable alleles was reported (Wang et al. 2012b; Zhang et al. 2012). We observed the same in linseed, an observation that should be carefully considered because the additive effects of the five QTL could be capitalized upon to directly improve TSW and indirectly yield. Interestingly, none of the modern linseed cultivars carried the five favorable alleles, indicating that further improvement of TSW within the modern linseed gene pool is feasible by MAS. The new Canadian cultivar AAC Bravo registered in 2012, possesses high TSW (6.8 g) that is well above the current Canadian linseed varieties ranging 5-5.5 g, and yields similar to CDC Bethune (Scott Duguid, personal communication, 2013). Independent marker testing of this variety that was not part of the core collection showed that it possesses four of the five TSW favorable alleles (data not shown). In addition to providing validation to our TSW markers, AAC Bravo illustrates a practical example of indirect yield improvement through yield components. However, additional validation in biparental populations testing various genetic backgrounds is warranted before implementation of molecular breeding strategies.

5.6 Conclusion

The current study provides initial insights into the genomic regions underlying agronomics traits. Although only 12 marker-trait associations were identified for six agronomic traits, these markers were consistent across environments and mostly stable. An attribute of AM is the identification and validation of favorable alleles in germplasm collections (Wang et al. 2012b). The accessions carrying favorable alleles, especially for TSW, will be useful to ensure their transfer into the best modern linseed cultivars. To further disentangle the genetic bases of yield and yield-related traits, marker density will be increased with thousands of single nucleotide polymorphism markers obtained by the re-sequencing of the entire core collection that should enable us to take advantage of the existing and comprehensive phenotypic data and the germplasm resources represented in the Canadian flax core collection (Diederichsen et al. 2013).

6.0 GENERAL DISCUSSION AND CONCLUSION

Flax (*Linum usitatissimum* L.), one of the founder crops of domesticated plants, was first described in Near Eastern agriculture around 9,000 BC. Despite the high global market prospects of linseed, it remains in Canada a secondary crop to wheat and canola. Canola has surpassed linseed acreage and yield, mainly as a consequence of the advent of hybrids and herbicide tolerant cultivars that facilitated its agronomic management and improved its performance. Conversely, linseed yields have recently plateaued, essentially because of the narrow genetic diversity used in the development of Canadian cultivars and the difficulties of accessing a wider variability without suitable genomic resources.

The comprehensive characterization of core collections is more amenable for breeding purposes because they allow accessing of the non-redundant variation present in large germplasm collections in a cost-effective manner. Similarly, the availability of genomic resources such as molecular markers, linkage and physical maps and whole genome sequencing are of paramount importance to fully exploit flax genetic resources. Genomic resources along with the Canadian flax core collection have been developed establishing the foundation for genomic assisted breeding for flax.

QTL mapping based on bi-parental crosses has been the most applied approach to map QTL associated with economic traits in crops. However, its limited allelic diversity per locus, its low mapping resolution and the time required to construct suitable populations, limit its usefulness as the only approach to accelerate linseed breeding in a timely manner. On the other hand, AM circumvents QTL mapping limitations, providing higher mapping resolution, the simultaneous identification of a larger number of

favorable alleles without requiring the construction of mapping populations. AM's main drawback is probably the poor power of detection of rare allele effects.

Here, we report on the genetic characterization of the Canadian flax core collection and its utilization for conducting AM for seven seed quality traits and nine agronomic traits evaluated at up to eight environments. The 407 accessions of the core collection were assigned to two major groups (G1 and G3) and six sub-groups, showing weak population structure, abundant genetic diversity and relatively fast genome-wide LD decay (~1 cM). For the seven seed quality traits, thirty-one stable candidate QTL were identified. Candidate QTL for PAL, LIO and LIN co-localized with QTL previously identified in bi-parental populations and some mapped nearby genes known to be involved in the FA biosynthesis pathway. Twelve consistent marker-trait associations were identified for the agronomic traits including five markers for TSW, one each for FL5%, FL95% and PB and two each for PH and LDG. Marker Lu943 showed associations with FL5%, FL95% and PH, and the five markers associated with TSW exhibited additive genetic effects accounting for 30% of the trait variation. Overall, the genetic characterization of the Canadian flax core collection and the candidate QTL and markers reported herein, stand for one of the largest flax genetic studies reported to date, and the first QTL study in flax using the AM approach, which represent the starting point for flax molecular breeding.

A suitable core collection for AM should encompass as much genetic diversity as possible, weak population and family structure and fast LD decay. Understanding the genetic structure of core collections is critical to control false positives in AM and the genetic diversity harbored by them has profound effects on LD decay and, consequently,

on mapping resolution. Combined population structure analysis grouped the core collection in two major groups mainly, but not exclusively, according to geographical origin, showing weak population subdivision (F_{ST} = 0.094). Within G1, the South Asian sub-group was the most genetically distinct. In line with previous reports, the North American sub-group of G3 reflects historical germplasm exchange between U.S.A. and Canada. More than 80% of the pairwise coancestry estimates among the accessions ranged from 0.1 to 0.3, indicating that most of the lines had weak family relatedness. An average of 5.32 alleles per locus was observed in the core collection, even exceeding that of a diverse sample of pale flax (4.62), the wild progenitor of cultivated flax. Interestingly, the Western European sub-group was remarkably rich in private alleles (246) despite its small population size (37). This sub-group could be a reservoir of novel genetic variation useful to expand the diversity of Canadian cultivars. This abundant genetic diversity was also reflected in the observed patterns of LD, with linked LD (LD caused by physical linkage) being predominant across major groups and sub-groups, declining relatively fast, a result generally unexpected for a self-pollinated species. Hence, our LD analysis highlights the importance of maximizing the genetic diversity when core collections or AM panels are assembled in order to enhance the probabilities of QTL discovery and the mapping resolution.

Linseed oil and its FA profile are the main factors determining its marketability. Nowadays, the industry's demand for novel products imposes a time constraint for breeders that is much faster than in the past. The key elements needed by linseed breeders to satisfy this demand include knowledge about the genetic diversity and the genetic bases of traits. This implies understanding the allelic diversity of QTL, the number of

QTL and their effects on traits, and the QTL by environment interaction patterns. We evaluated 390 accessions from the core collection across six environments and phenotypic data was collected for OIL, PAL, STE, OLE, LIO, LIN and IOD. Significant GE interaction was observed for all seven seed quality traits, implying that specific enzymes involved in the FA biosynthesis pathway were more sensitive to differences in temperature, precipitation, soil conditions, etc. The traits showed broad phenotypic variation involving a significant contribution from genetic factors. Model comparison showed that the MLM (PCA + *K*) was the best to conduct the AM analysis for seed quality traits because minimized the number of false positive associations caused by population and family structures. To date, only three FA-related QTL have been identified in flax. In the present study, one of them was validated, i.e., the co-located *QLio-LG12.3* and *QLin-LG12.3* and several new markers and candidate QTL were mapped nearby genes involved in the FA biosynthesis pathway. Marker Lu566 (LG7) associated with LIO and LIN mapped close to the *fad3A* gene, hence making it a candidate gene for the QTL underlying these traits. Further, the validated QTL intervals were noticeably narrowed down compared with previous reports, making them more amenable for MAS. However, large differences in the variation accounted for the QTL reported herein, and those identified by biparental mapping were observed. Low marker coverage, incomplete LD, removal of rare alleles and epigenetic variations are considered potential factors underpinning these discrepancies. Nevertheless, various major QTL explained up to 19% of the phenotypic variation, suitable for MAS. Canadian cultivars displayed eight times narrower diversity compared with the core collection as revealed by the comparison of frequency of QTL/marker alleles, indicating that further genetic

improvement could be made for seed quality traits. Overall, the candidate QTL and markers identified in this study provide the initial insights into the genetic architecture of seed quality traits and will establish the foundation for future marker assisted breeding in linseed.

The combination of a suitable FA profile and agronomic performance can readily increase linseed competitiveness. Seed quality without agronomic fitness poses a threat to farmers' profitability. As a consequence, breeding effort based on quality only, has little effect on increasing linseed acreage with no chance of stopping the vicious circle. To tackle this problem, yield, yield components and other correlated agronomic traits can be bred simultaneously, because generally phenotypic correlations point to genetic linkage or pleiotropy. Thus, an understanding of these traits is of practical value to breeders because such information assists in the design of efficient breeding strategies.

To contribute to linseed competitiveness, the Canadian flax core collection was phenotyped for nine agronomic traits including yield, BPA, TSW, SPB, FL5%, FL95%, PH, PB and LDG over eight environments. Broad phenotypic variation was observed with significant contribution accounted for by the genotypes (G) and the GE interactions, which highlights the necessity of pinpointing the most stable genotypes harbouring favorable alleles. Significant correlations between yield and its components, FL5%, FL95%, PH and LDG, denoted that further yield and overall agronomic performance can be improved by exploiting these correlations.

Flax morphotypes and geographic origin were the main factors determining the population structure observed in the flax core collection as revealed by the structure analysis carried out within the major groups G1 and G3. Although the effect of

population structure on trait variation was only significant for PH, we cannot discard the potential effect of the fibre morphotype on yield and its components because it is likely that the favorable alleles associated with these traits do not segregate homogeneously across groups, or they could even be totally absent in the fibre accessions which have not been selected for them, consequently under powering the AM results. AM analysis without the fibre accessions could provide further insights in this regard.

The number of marker-trait associations was significantly affected by environmental factors. The environment specific patterns observed for some significant associations support the need of combining phenotypic selection across multiple environments with MAS to increase QTL selection accuracy. Markers such as Lu943 associated with FL5%, FL95% and PH, and Lu2532 and Lu526 associated with TSW exhibited the largest phenotypic effects and high stability illustrating their potential for improving yield through yield components and the overall agronomic fitness of linseed cultivars. None of the flax accessions and modern cultivars carried all of the five favorable alleles associated with TSW. Nevertheless, genotypes carrying four alleles were identified, which should help in parental selection to speed up allele pyramiding in adapted genetic background.

Our genetic analyses confirmed that the Canadian flax core collection, comprising variation from 76 countries, represents the major reservoir of genetic diversity for this species reported to date. Its allelic richness has the potential to contribute significantly to flax molecular breeding as demonstrated through AM analysis. To realize the full potential of AM and the flax core collection, marker density will be increased with thousands of single nucleotide polymorphism markers obtained by the re-sequencing of

the entire core collection. This resource should enable us to take advantage of the existing and comprehensive phenotypic data generated from the characterization of the Canadian flax core collection. Further validation of the QTL reported herein in biparental population as well as studies of the potential adverse effects of the fibre morphotype on the number of marker-trait associations are required.

7.0 LITERATURE CITED

Abdurakhmonov I, Abdukarimov A (2008) Application of association mapping to understanding the genetic diversity of plant germplasm resources. Int J Plant Genomics 2008:574927

Abdurakhmonov IY, Kohel RJ, Yu JZ, Pepper AE, Abdullaev AA, Kushanov FN, Salakhutdinov IB, Buriev ZT, Saha S, Scheffler BE, Jenkins JN, Abdukarimov A (2008) Molecular diversity and association mapping of fiber quality traits in exotic *G. Hirsutum* L. germplasm. Genomics 92:478-487

Abdurakhmonov IY, Saha S, Jenkins JN, Buriev ZT, Shermatov SE, Scheffler BE, Pepper AE, Yu JZ, Kohel RJ, Abdukarimov A (2009) Linkage disequilibrium based association mapping of fiber quality traits in *G. hirsutum* L. variety germplasm. Genetica 136:401-417

Achleitner A, Tinker N, Zechner E, Buerstmayr H (2008) Genetic diversity among oat varieties of worldwide origin and associations of AFLP markers with quantitative traits. Theor Appl Genet 117:1041-1053

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. BMC Bioinformatics 9:323

Aranzana MJ, Kim S, Zhao K, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang C, Toomajian C, Traw B, Zheng H, Bergelson J, Dean C, Marjoram P, Nordborg M (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet 1:e60

Ardlie K, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3:299-309

Arnheim N, Calabrese P, Nordborg M (2003) Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. Am J Hum Genet 73:5-16

Arsovski AA, Villota MM, Rowland O, Subramaniam R, Western TL (2009) *MUM ENHANCERS* are important for seed coat mucilage production and mucilage secretory cell differentiation in *Arabidopsis thaliana*. J Exp Bot 60:2601-2612

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology tool for the unification of biology. Nat Genet 25:25-29

Ashikari M, Sakakibara H, Lin S, Yamamoto T, Takashi T, Nishimura A, Angeles ER, Qian Q, Kitano H, Matsuoka M (2005) Cytokinin oxidase regulates rice grain production. Science 309:741-745

Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annu Rev Plant Genet 44:293-308

Avachat AM, Dash RR, Shrotriya SN (2011) Recent investigations of plant based natural gums, mucilages and resins in novel drug delivery systems. Ind J Pharm Edu Res 45:86-99

Bachlava E, Dewey RE, Burton JW, Cardinal AJ (2009) Mapping candidate genes for oleate biosynthesis and their association with unsaturated fatty acid seed content in soybean. Mol Breed 23:337-347

Balding D (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7:781-791

Banik M, Duguid S, Cloutier S (2011) Transcript profiling and gene characterization of three fatty acid desaturase genes in high, moderate, and low linolenic acid genotypes of flax (*Linum usitatissimum* L.) and their role in linolenic acid accumulation. Genome 54:471-83

Bansal V, Libiger O, Torkamani A, Schork N (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11:773-785

Barbary OM, Al-Sohaimy SA, El-Saadani MA, Zeitoun AMA (2009) Extraction, composition and physicochemical properties of flaxseed mucilage. J Adv Agric Res 14:605-622

Bar-Hen A, Charcosset A, Bourgoin M, Guiard J (1995) Relationship between genetic markers and morphological traits in a maize inbred line collection. Euphytica 84:145-154

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proc Royal Soc Lond B 263:1619-1626

Beer S, Siripoonwiwat W, O'Donoughue L, Souza E, Mathews D, Sorrells M (1997) Associations between molecular markers and quantitative traits in oat germplasm pool: can we infer linkages? J Agric Genom 3

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. J R Statist Soc B 57:289-300

Bernardo R, Romero-Severson J, Ziegle J, Hauser J, Joe L, Hookstra G, Doerge RW (2000) Parental contribution and coefficient of coancestry among maize inbreds: pedigree, RFLP, and SSR data. Theor Appl Genet 100:552-556

Bhatty RS (1993) Further compositional analyses of flax: mucilage, trypsin inhibitors and hydrocyanic acid. J Am Oil Chem Soc 70:899-904

Bickel CL, Gadani S, Lukacs M, Cullis CA (2011) SSR markers developed for genetic mapping in flax (*Linum usitatissimum* L.). Res Rep Biol 2:23-29

Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, Bergelson J, Cuguen J, Roux F (2010) Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. PLoS Genet 6:e1000940

Bradbury PJ, Parker T, Hamblin MT, Jannink JL (2011) Assesment of power and false discovery rate in genome-wide association studies using the BarleyCAP germplasm. Crop Sci 51:52-59

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635

Breseghello F, Sorrells M (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165-1177

Caballero A, Toro MA (2002) Analysis of genetic diversity for the management of conserved subdivided populations. Conserv Genet 3:289-299

Cadic E, Coque M, Vear F, Grezes-Besset B, Pauquet J, Piquemal J, Lippi Y, Blanchard P, Romestant M, Pouilly N, Rengel D, Gouzy J, Langlade N, Mangin B, Vincourt P

(2013) Combined linkage and association mapping of flowering time in sunflower (*Helianthus annuus* L.). Theor Appl Genet 126:1337-56

Camus-Kulandaivelu L, Veyrieras JB, Madur D, Combes V, Fourmann M, Barraud S, Dubreuil P, Gouesnard B, Manicacci D, Charcosset A (2006) Maize adaptation to temperate climate: relationship between population structure and polymorphism in the *Dwarf8* gene. Genetics 172:2449-2463

Caniato FF, Guimarães CT, Hamblin M, Billot C, Rami JF, Hufnagel B, Kochian LV, Liu J, Garcia AA, Hash CT, Ramu P, Mitchell S, Kresovich S, Oliveira AC, de Avellar G, Borém A, Glaszmann JC, Schaffert RE, Magalhaes JV (2011) The relationship between population structure and aluminum tolerance in cultivated sorghum. PLoS One 6:e20830

Casa AM, Mitchell SE, Hamblin MT, Sun H, Bowers JE, Paterson AH, Aquadro CF, Kresovich S (2005) Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. Theor Appl Genet 111:23-30

Casa R, Russell G, Lo Cascio B, Rossini F (1999) Environmental effects on linseed (*Linum usitatissimum* L.) yield and growth of flax at different stand densities. Eur J Agron 11:267-278

Chai Y, Hao X, Yang X, Allen WB, Li J, Yan J, Shen B, Li J (2012) Validation of *DGAT1-2* polymorphisms associated with oil content and development of functional markers for molecular breeding of high-oil maize. Mol Breed 29:939-949

Chen Y, Dribnenki P (2004) Effect of medium osmotic potential on callus induction and shoot regeneration in flax anther culture. Plant Cell Rep 23:272-276

Chen Y, Lubberstedt T (2010) Molecular basis of trait correlations. Trends Plant Sci 15:454-461

Chung J, Babka HL, Graef GL, Staswick PE, Lee DJ, Cregan PB, Shoemaker RC, Specht JE (2003) The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci 43:1053-1067

Clark RM, Wagler TN, Quijada P, Doebley J (2006) A distant upstream enhancer at the maize domestication gene *tb1* has pleiotropic effects on plant and inflorescent architecture. Nat Genet 38:594-597

Cloutier S, Miranda E, Ward K, Radovanovic N, Reimer E, Walichnowski A, Datla R, Rowland G, Duguid S, Ragupathy R (2012a) Simple sequence repeat marker development from bacterial artificial chromosome end sequences and expressed sequence tags of flax (*Linum usitatissimum* L.). Theor Appl Genet 125:685-694

Cloutier S, Niu Z, Datla R, Duguid S (2009) Development and analysis of EST-SSRs for flax (*Linum usitatissimum* L.). Theor Appl Genet 119:53-63

Cloutier S, Ragupathy R, Niu Z, Duguid S (2011) SSR-based linkage map of flax (*Linum usitatissimum* L.) and mapping of QTLs underlying fatty acid composition traits. Mol Breed 28:437-451

Cloutier S, Ragupathy R, Miranda E, Radovanovic N, Reimer E, Walichnowski A, Ward K, Rowland G, Duguid S, Banik M (2012b) Integrated consensus genetic and physical maps of flax (*Linum usitatissimum* L.). Theor Appl Genet 125:1783-1795

Cockram J, Jones H, Leigh FJ, O'Sullivan D, Powell W, Laurie DA, Greenland AJ (2007) Control of flowering time in temperate cereals: genes, domestication, and sustainable productivity. J Exp Bot 58:1231-1244

Cockram J, White J, Zuluaga DL, Smith D, Comadran J, Macaulay M, Luo Z, Kearsey MJ, Werner P, Harrap D, Tapsell C, Liu H, Hedley PE, Stein N, Schulte D, Steuernagel B, Marshall DF, Thomas WT, Ramsay L, Mackay I, Balding DJ; AGOUEB Consortium, Waugh R, O'Sullivan DM (2010) Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. Proc Natl Acad Sci USA 107:21611-21616

Cook JP, McMullen MD, Holland JB, Tian F, Bradbury P, Ross-Ibarra J, Buckler ES, Flint-Garcia SA (2012) Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant Physiol 158:824-34

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics 144:2001-2014

Cullis C (2011) Linum. In: Kole C (ed) Wild crop relatives: genomic and breeding resources oilseeds. Springer, New York, pp 177-189

Cullis CA (2007) Flax. In: Kole C (ed) Genome mapping and molecular breeding in plants, vol 2. Springer, Berlin, pp 275-295

Czemplik M, Boba A, Kostyn K, Kulma A, Mituła A, Sztajnert M, Wróbel-Kwiatkowska M, Żuk M, Jan Szopa J, Skórkowska-Telichowska K (2011) Flax engineering for biomedical application. In: Komorowska MA, Olsztynska-Janus S (eds) Biomedical engineering, trends, research and technologies. InTech, Rijeka, pp 407-434

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganske R, Lõhmussaar E, Zernant J, Tõnisson N, Remm M, Mägi R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P, Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I

(2002) A first-generation linkage disequilibrium map of human chromosome 22. Nature 418:544-548

Dean GH, Zheng H, Tewari J, Huang J, Young DS, Hwang YT, Western TL, Carpita NC, McCann MC, Mansfield SD, Haughn GW (2007) The Arabidopsis *MUM2* gene encode a β-galactosidase required for the production of seed coat mucilage with correct hydration properties. Plant Cell 19:4007-4021

Deng X, Long S, He D, Li X, Wang Y, Hao D, Qiu C, Chen X (2011) Isolation and characterization of polymorphic microsatellite markers from flax (*Linum usitatissimum* L.). Afr J Biotechnol 10:734-739

Deng X, Long S, He D, Li X, Wang Y, Liu J, Chen H (2010) Development and characterization of polymorphic microsatellite markers in *Linum usitatissimum*. J Plant Res 123:119-123

Diederichsen A (2001) Comparison of genetic diversity of flax (*Linum usitatissimum* L.) between Canadian cultivars and a world collection. Plant Breed 120:360-362

Diederichsen A (2007) Ex situ collections of cultivated flax (*Linum usitatissimum* L.) and other species of the genus Linum L. Genet Resour Crop Evol 54:661-678

Diederichsen A, Fu YB (2006) Phenotypic and molecular (RAPD) differentiation of four infraspecific groups of cultivated flax (*Linum usitatissimum* L. subp. *usitatissimum*). Genet Resour Crop Evol 53:77-90

Diederichsen A, Fu BF (2008) Flax genetic diversity as the raw material for future success. International conference on flax and other bast plants [http:// www.saskflax.com/documents/fb_papers/51_Diederichsen.pdf]

Diederichsen A, Hammer K (1995) Variation of cultivated flax (*Linum usitatissimum* L. subp. u*sitatissimum*) and its wild progenitor pale flax (subsp. *angustifolium* (Huds.) Thell.). Genet Resour Crop Evol 42:263-272

Diederichsen A, Raney JP (2006) Seed colour, seed weight and seed oil content in *Linum usitatissimum* accessions held by Plant Gene Resources of Canada. Plant Breed 125:372-377

Diederichsen A, Richards K (2003) Cultivated flax and the genus *Linum* L., taxonomy and germplasm conservation. In: Muir AD, Westcott ND (eds) Flax the genus *Linum*. CRC Press, Boca Raton, pp 22-54

Diederichsen A, Ulrich A (2009) Variability in stem fibre content and its association with other characteristics in 1177 flax (*Linum usitatissimum* L) genebank accessions. Ind Crop Prod 30:33-39

Diederichsen A, Kusters PM, Kessler D, Bainas Z, Gugel RK (2013) Assembling a core collection from the flax world collection maintained by Plant Gene Resources of Canada. Genet Resour Crop Evol 60:1479-1485

Diederichsen A, Rozhmina TA, Zhuchenko A, Richards K (2006) Screening for broad adaptation in 96 flax (*Linum usitatissimum* L.) accessions under dry and warm conditions in Canada and Russia. Plant Genet Resour Newsl 14:9-16

Du F, Clutter A, Lohuis M (2007) Characterizing linkage disequilibrium in pig populations. Int J Biol Sci 3:166-178

Duguid SD (2009) Flax. In: Vollmann J, Rajcan I (eds) Oil crops, Handbook of plant breeding 4. Springer, New York, pp 233-255

Dunning AM, Durocher F, Healey CS, Teare MD, McBride SE, Carlomagno F, Xu CF, Dawson E, Rhodes S, Ueda S, Lai E, Luben RN, Van Rensburg EJ, Mannermaa A, Kataja V, Rennart G, Dunham I, Purvis I, Easton D, Ponder BAJ (2000) The extent of linkage disequilibrium in four populations with distinct demographic histories. Am J Hum Genet 67:1544–54

Ersoz E, Yu J, Buckle, E (2007) Application of linkage disequilibrium and association mapping in crop plants. In: Varshney R, Tuberosa R (eds) Genomics-assisted crop improvement. Springer, Dordrecht, pp 97-119

Ersoz ES, Yu J, Buckler ES (2009) Applications of linkage disequilibrium and association mapping in maize. In: Kriz A, Larkins B (eds) Molecular genetic approaches to maize improvement. Springer, Berlin, pp 173-195

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611-2620

Everaert I, De Riek J, De Loose M, Van Waes J, Van Bockstaele E (2001) Most similar variety grouping for distinctness evaluation of flax and linseed (*Linum usitatissimum* L.) varieties by means of AFLP and morphological data. Plant Var Seed 14:69-87

Excoffier L, Heckel G (2006) Computer programs for population genetics data analysis: a survival guide. Nat Rev Genet 7:745-758

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity 103:285-298

Excoffier L, Lischer HEL (2010) Arlequin suite ver. 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564-567

FAOSTAT (2013) Production of crops: linseed: area harvested and production (tonnes). Available at http://faostat3.fao.org/home/index.html Accessed March 2013

Fenart S, Ndong YP, Duarte J, Rivière N, Wilmer J, van Wuytswinkel O, Lucau A, Cariou E, Neutelings G, Gutierrez L, Chabbert B, Guillot X, Tavernier R, Hawkins S, Thomasset B (2010) Development and validation of a flax (*Linum usitatissimum* L.) gene expression oligo microarray. BMC Genomics 11:592

Fisher RA (1935) The logic of inductive inference. J R Stat Soc Ser A 98:39-54

Flint-Garcia S, Thornsberry JM, Bukler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357-374

Flori L, Fritz S, Jaffrézic F, Boussaha M, Gut I, Heath S, Foulley JL, Gautier M (2009) The genome response to artificial selection: a case study in dairy cattle. PLoS ONE 4:e6595

Fofana B, Cloutier S, Duguid S, Ching J, Rampitsch C (2006) Gene expression of stearoyl-ACP desaturase and Δ12 fatty acid desaturase 2 is modulated during seed development of flax (*Linum usitatissimum*). Lipids 41:705-720

Fofana B, Duguid S, Cloutier S (2004) Cloning of fatty acid biosynthetic genes *β*-ketoacyl CoA synthase, fatty acid elongase, stearoyl-ACP desaturase and fatty acid desaturase and analysis of expression in the early developmental stages of flax (*Linum usitatissimum* L.) seeds. Plant Sci 166:1487-1496

Friedt W, Bickert C, Schaub H (1995) In vitro breeding of high linolenic, doubled haploid lines of linseed (*Linum usitatissimum* L.) via androgenesis. Plant Breed 114:322-326

Fu YB (2005) Geographic patterns of RAPD variation in cultivated flax. Crop Sci 45:1084-1091

Fu YB (2011) Genetic evidence for early flax domestication with capsular dehiscence. Genet Resour Crop Evol 58:1119-1128

Fu YB, Diederichsen A, Richards KW, Peterson G (2002) Genetic diversity within a range of cultivars and landraces of flax (*Linum usitatissimum* L) as revealed by RAPDs. Genet Resour Crop Evol 49:167-174

Fu YB, Rowland GG, Duguid SD, Richards K (2003) RAPD analysis of 54 North American flax cultivars. Crop Sci 43:1510-1515

Garris A, Tai T, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631-1638

Gauch HG (1992) AMMI analysis of yield trials. In: Kang MS, Gauch HG (eds) Genotype-by-environment interaction. CRC Press, Boca Raton, pp 1-40

Gaut B, Long A (2003) The lowdown of linkage disequilibrium. Plant Cell 15:1502-1506

Gill KS, Yermanos DM (1967) Cytogenetic studies on the genus *Linum* I. Hybrids among taxa with 15 as the haploid chromosome number. Crop Sci 7:623-627

Goldman IL, Rocheford TR, Dudley JW (1994) Molecular markers associated with maize kernel oil concentration in an Illinois high protein x Illinois low protein cross. Crop Sci 34:908-915

Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82:100-112

Green AG (1986) A mutant genotype of flax (*Linum usitatissimum* L.) containing very low levels of linolenic acid in its seed oil. Can J Plant Sci 66:499-503

Green AG, Marshall AR (1984) Isolation of induced mutants in linseed (*Linum usitatissimum* L.) having reduced linolenic acid content. Euphytica 33:321-328

Green AG, Chen Y, Singh SP, Dribnenki JCP (2008) Flax. In: Kole C, Hall TC (eds) Compendium of transgenic crop plants: Transgenic oilseed crops. Blackwell Publishing , Chicester, pp 199-226

Guilloux K, Gaillard I, Courtois J, Courtois B, Petit E (2009) Production of arabinoxylan-oligosaccharides from flaxseed (*Linum usitatissimum*). J Agric Food Chem 57:11308-11313

Guo B, Wang D, Guo Z, Beavis WD (2013) Family-based association mapping in crop species. Theor Appl Genet 126:1419-1430

Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol 57:461-85

Gutierréz JP, Royo LJ, Álvarez I, Goyache F (2005) MolKin v.2.0 A computer program for genetic analysis of populations using molecular coancentry information. J Hered 96:718-721

Hamdan YAS, Garcia-Moreno MJ, Fernandez-Martinez JM, Velasco L, Perez-Vich B (2012) Mapping of major and modifying genes for high oleic acid content in safflower. Mol Breed 30:1279-1293

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2:618-620

Harris DR (1997) The spread of neolithic agriculture from the Levant to western-central Asia. In: Damania AB, Valkoun J, Willcox G, Qualset CO (eds) The origin of agriculture and crop domestication. Proc Harlan Symp, ICARDA, Aleppo, Syria, May 10-14, pp 65-82

He XC, Qin YM, Xu Y, Hu CY, Zhu YX (2008) Molecular cloning, expression profiling, and yeast complementation of 19 β-tubulin cDNAs from developing cotton ovules. J Exp Bot 59:2687-2695

Hill W, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38: 226-231

Holland JB (2007) Genetic architecture of complex traits in plants. Curr Opin Plant Biol 10:156-161

Honsdorf N, Becker HC, Ecke W (2010) Association mapping for phenological, morphological, and quality traits in canola quality winter rapeseed (*Brassica napus* L.). Genome 53:899-907

Hu X, Sullivan-Gilbert M, Gupta M, Thompson SA (2006) Mapping of the loci controlling oleic and linolenic acid contents and development of *fad2* and *fad3* allele-specific markers in canola (*Brassica napus* L.). Theor Appl Genet 113:497-507

Huang J, Zhang J, Li W, Hu W, Duan L, Feng Y, Qiu F, Yue B (2013) Genome-wide association analysis of ten chilling tolerance indices at the germination and seedling stages in maize. J Integr Plant Biol 55:735-744

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T,

Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang QF, Li J, Han B (2010a) Genome-wide studies of 14 agronomic traits in rice landraces. Nat Genet 42:961-967

Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, Li W, Guo Y, Deng L, Zhu C, Fan D, Lu Y, Weng Q, Liu K, Zhou T, Jing Y, Si L, Dong G, Huang T, Lu T, Feng Q, Qian Q, Li J, Han B (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 44:32-39

Huang YF, Madur D, Combes V, Ky CL, Coubriche D, Jamin P, Jouanne S, Dumas F, Bouty E, Bertin P, Charcosset A, Moreau L (2010b) The genetic architecture of grain yield and related traits in *Zea maize* L. revealed by comparing intermated and conventional populations. Genetics 186:395-404

Hubisz M, Falush D, Stephens M, Pritchard J (2009) Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9:1322-1332

Ishimaru K (2003) Identification of a locus increasing rice yield and physiological analysis of its function. Plant Physiol 133:1083-1090

Jarillo JA, Piñeiro M (2011) Timing is everything in plant development. The central role of floral repressors. Plant Sci 181:364-378

Jennings H (1917) The numerical results of diverse systems of breeding, with respect to two pairs of characters, linked or independent, with special relation to the effects of linkage. Genetics 2:97-154

Jhala AJ, Weselake RJ, Hall LM (2009) Genetically engineered flax: Potential benefits, risks, regulations, and mitigation of transgene movement. Crop Sci 49:1943-1954

Jordan MC, McHughen A (1988) Glyphosate tolerant flax plants from Agrobacterium mediated gene-transfer. Plant Cell Rep 7:281-284

Jung M, Ching A, Bhattramakki D, Dolan M, Tingey S, Morgante M, Rafalski A (2004) Linkage disequilibrium and sequence diversity in a 500-kbp region around the *adh*1 locus in elite maize germplasm. Theor Appl Genet 109:681-689

Kale SM, Pardeshi VC, Kadoo NY, Ghorpade PB, Jana MM, Gupta VS (2012) Development of genomic simple sequence repeat markers for linseed using next-generation sequencing technology. Mol Breed 30:597-606

Kalinowski ST (2005) HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness. Mol Ecol Notes 5:187-189

Kauer MO, Dieringer D, Schlötterer C (2003) A microsatellite variability screen for positive selection associated with the "out of Africa" habitat expansion of *Drosophila melanogaster*. Genetics 165:1137-1148

Kenaschuk EO (2005) High linolenic acid flax. US patent 6870077 issued on March 22, 2005

Khadake RM, Ranjekar PK, Harsulkar AM (2009) Cloning of a novel omega-6 desaturase from flax (*Linum usitatissimum*) and its functional analysis in *Saccharomyces cerevisiae*. Mol Biotechnol 42:168-174

Kovach MJ, Sweeney MT, McCouch SR (2007) New insights into the history of rice domestication. Trends Genet 23:578-587

Krill A, Kirst M, Kochian L, Buckler E, Hoekenga O (2010) Association and linkage analysis of aluminum tolerance genes in maize. PLoS One 5:e9958

Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. J Amer Statist Assoc 47:583-621

Kumar S, You FM, Cloutier S (2012) Genome wide SNP discovery in flax through next generation sequencing of reduced representation libraries. BMC Genomics 13:684

Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, Oropeza-Rosas MA, Zwonitzer JC, Kresovich S, McMullen MD, Ware D, Balint-Kurti PJ, Holland JB (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat Genet 43:163-168

Lander ES, Botstein DB (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199

Lewontin C (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49-67

Li B, Leal S (2008) Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. Am J Hum Genet 83:311-321

Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, Han Y, Chai Y, Guo T, Yang N, Liu J, Warburton ML, Cheng Y, Hao X, Zhang P, Zhao J, Liu Y, Wang G, Li J, Yan J (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet 45:43-50

Li X, Wei Y, Moore K, Michaud R, Viands D, Hansen J, Acharya A, Brummer EC (2011a) Association mapping of biomass yield and stem composition in a tetraploid alfalfa breeding population. Plant Genome 4:24-35

Li X, Yan W, Agrama H, Jia L, Jackson A, Moldenhauer K, Yeater K, McClung A, Wu D (2012) Unraveling the complex trait of harvest index with association mapping in rice (*Oryza sativa* L.). PLoS One 7:e29350

Li X, Yan W, Agrama H, Jia L, Shen X, Jackson A, Moldenhauer K, Yeater K, McClung A, Wu D (2011b) Mapping QTLs for improving grain yield using the USDA rice mini-core collection. Planta 234:347-61

Li Y, Haseneyer G, Schön C, Ankerst D, Korzun V, Wilde P, Vauer E (2011c) High levels of nucleotide diversity and fast decline of linkage disequilibrium in rye (*Secale cereale* L.) genes involved in frost response. BMC Plant Biol 11:1-14

Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz J (2010) Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 107:21119-21204

Li Y, Smulders MJM, Chang R, Qiu L (2011d) Genetic diversity and association mapping in a collection of selected Chinese soybean accessions based on SSR marker analysis. Conserv Genet 12:1145-1157

Lin CS, Poushinsky G (1985) A modified augmented design (type 2) for rectangular plots. Can J Plant Sci 65:743-749

Lin J, Quinn TP, Hilborn R, Hauser L (2008) Fine-scale differentiation between sockeye salmon ecotypes and the effect of phenotype on straying. Heredity 101:341-350

Liu A, Burke JM (2006) Patterns of nucleotide diversity in wild and cultivated sunflowers. Genetics 173:321-330

Liu K, Muse S (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128-2129

Liu W, Kim MY, Van K, Lee YH, Li H, Liu X, Lee SH (2011) QTL identification of yield-related traits and their association with flowering and maturity in soybean. J Crop Sci Biotech 14:65-70

Lu Y, Zhang S, Shah T, Xie C, Hao Z, Li X, Farkhari M, Ribaut JM, Cao M, Rong T, Xu Y (2010) Joint linkage-linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. Proc Natl Acad Sci USA 107:19585-19590

Luikart G, Allendorf FW, Cornuet JM, Sherwin WB (1998) Distortion of allele frequency distributions provides a test for recent population bottlenecks. J Hered 89:238–247

Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends Plant Sci 12:57-63

Maggioni L, Pavelek M, van Soest LJM, Lipman E (2002) Flax genetic resources in Europe. Ad hoc meeting: 7-8 December 2001; Prague. Czech Republic: International Plant Genetic Resources Institute

Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. Bioinformatics 20:2878-2879

Malysheva-Otto L, Ganal M, Röder M (2006) Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). BMC Genetics 7:1-14

Manly BFJ (1985) The statistics of natural selection. In: Chapman and Hall (ed) Spurious test results due to isolation by distance. Chapman and Hall, London, pp 186-195

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747-53

Månsby E, Diaz O, von Bothmer R (2000) Preliminary study of genetic diversity in Swedish flax (*Linum usitatissimum*). Genet Resour Crop Evol 47:417-424

McCarthy FM, Wang N, Magee GB, Nanduri B, Lawrence ML, Camon EB, Barrell DG, Hill DP, Dolan ME, Williams WP, Luthe DS, Bridges SM, Burgess SC (2006) AgBase: a functional genomics resource for agriculture. BMC Genomics 7:229

McHughen A, Swartz M (1984) A tissue-culture derived salt-tolerant line of flax (*Linum usitatissimum*). J Plant Physiol 117:109-117

Melchinger AE, Utz HF, Schön CC (2004) QTL analyses of complex traits with cross validation, bootstrapping and other biometric methods. Euphytica 137:1-11

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829

Morrell PL, Toleno DM, Lundy KE, Clegg MT (2005) Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization. Proc Natl Acad Sci USA 102:2442-2447

Muravenko OV, Bolsheva NL, Yurkevich OYu, Nosova IV, Rachinskaya OA, Samatadze TE, Zelenin AV (2010) Karyogenomics of species of the genus *Linum* L. Russ J Genet 46:1182-1185

Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia JM, Ware D, Bustamante CD, Buckler ES (2011) Genetic structure and domestication history of the grape. Proc Natl Acad Sci USA 108:3530-3535

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Bukler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell 21:2194-2202

Naran R, Chen G, Carpita NC (2008) Novel rhamnogalacturonan I and arabinoxylan polysaccharides of flax seed mucilage. Plant Physiol 148:132-141

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321-3323

Nemri A, Atwell S, Tarone AM, Huang YS, Zhao K, Studholme DJ, Nordborg M, Jones JD (2010) Genome-wide survey of Arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping. Proc Natl Acad Sci USA 107:10302-10307

Nielsen D, Zaykin D (2001) Association mapping: where we′ve been, where we′re going. Expert Rev Mol Diagn 1:334-342

Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Pääbo S, Pritchard JK, Rubin EM (2006) Sequencing and analysis of Neanderthal genomic DNA. Science 314:1113-1118

Nordborg M (2000) Linkage disequilibrium, genes trees and selfing: An ancestral recombination graph with partial self-fertilization. Genetics 154:923-929

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, Stahl EA, Weigel D (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 30:190-193

Oh TJ, Gorman M, Cullis CA (2000) RFLP and RAPD mapping in flax (*L. usitatissimum*). Theor Appl Genet 101:590- 593

Oomah BD (2001) Flaxseed as a functional food source. J Sci Food Agric 81:889-894

Oomah BD, Mazza G (1993) Flaxseed proteins review. Food Chem 48:109-114

Oomah BD, Kenaschuk EO, Cui W, Mazza G (1995) Variation in the composition of water-soluble polysaccharides in flax seed. J Agric Food Chem 43:1484-1488

Oraguzie N, Wilcox P, Rikkerink H, de Silva H (2007) Linkage disequilibrium. In: Oraguzie NC, Rikkerink EHA, Gardiner SE, de Silve HN (eds) Association mapping in plants. Springer, New York, pp 11-39

Palaisa K, Morgante M, Tingey S, Rafalski A (2003) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell 15:1795-1806

Panthee DR, Pantalone VR, Saxton AM, West DR, Sams CE (2007) Quantitative trait loci for agronomic traits in soybean. Plant Breed 126:51-57

Parry MAJ, Hawkesford MJ (2012) An integrated approach to crop genetic improvement. J Integr Plant Biol 54:250-259

Pasam RK, Sharma R, Malosetti M, van Eeuwijk F, Haseneyer G, Kilian B, Graner A (2012) Genome-wide association studies for agronomical traits in a worldwide spring barley collection. BMC Plant Biol 12:16

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in excel. Population genetic software for teaching and research. Mol Ecol Notes 6:288-295

Peng B, Li Y, Wang Y, Liu C, Liu Z, Tan W, Zhang Y, Wang D, Shi Y, Sun B, Song Y, Wang T, Li Y (2011) QTL analysis for yield components and kernel-related traits in maize across multi-environments. Theor Appl Genet 122:1305-1320

Poland J, Bradbury P, Buckler E, Nelson R (2011) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. Proc Natl Acad Sci USA 108:6893-6898

Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. Trends Plant Sci 1:215-222

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904-909

Pritchard J (2001) Deconstructing maize population structure. Nat Genet 28:203-204

Pritchard J, Rosenberg N (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220-228

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170-181

Przybylski R (2001) Flax oil and high linolenic oils. In: Shahidi F (ed) Bailey's industrial oil and fat products. John Wiley and Sons Inc, Hoboken, pp 281-301

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575

Purchase JL (1997) Parametric analysis to describe G x E interaction and yield stability in winter wheat. PhD thesis. Department of Agronomy, Faculty of Agriculture, University of the Orange Free State, Bloemfontein, South Africa

Qi Z, Wu Q, Han X, Sun Y, Du X, Liu C, Jiang H, Hu G, Chen Q (2011) Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. Euphytica 179:499-514

Qiu D, Morgan C, Shi J, Long Y, Liu J, Li R, Zhuang X, Wang Y, Tan X, Dietrich E, Weihmann T, Everett C, Vanstraelen S, Beckett P, Fraser F, Trick M, Barnes S,

Wilmer J, Schmidt R, Li J, Li D, Meng J, Bancroft I (2006) A comparative linkage map of oilseed rape and its use for QTL analysis of seed oil and erucic acid content. Theor Appl Genet 114:67-80

Rachinskaya OA, Lemesh VA, Muravenko OV, Yurkevich OY, Guzenko EV, Bol'sheva NL, Bogdanova MV, Samatadze TE, Popov KV, Malyshev SV, Shostak NG, Heller K, Hotyleva LV, Zelenin AV (2011) Genetic polymorphism of flax *Linum usitatissimum* based on the use of molecular cytogenetic markers. Genetika 47:56-65

Rafalski JA (2010) Association genetics in crop improvement. Curr Opin Plant Biol 13:174-80

Ragupathy R, Rathinavelu R, Cloutier S (2011) Physical mapping and BAC-end sequence analysis provide initial insights into the flax (*Linum usitatissimum* L.) genome. BMC Genomics 12:217

Rajhathy T (1976) Haploid flax revisited. Z Pflanzeniichtg 76:1-10

Rajwade AV, Arora RS, Kadoo NY, Harsulkar AM, Ghorpade PB, Gupta VS (2010) Relatedness of Indian flax genotypes (*Linum usitatissimum* L.): an inter-simple sequence repeat (ISSR) primer assay. Mol Biotechnol 45:161-170

Rakyan VK, Down TA, Balding DJ, Beck S (2011) Epigenome-wide association studies for common human diseases. Nat Rev Genet 12:529-41

Raman H, Stodart B, Ryan PR, Delhaize E, Emebiri L, Raman R, Coombes N, Milgate A (2010) A genome-wide association analysis of common wheat (*Triticum aestivum* L.) germplasm identifies multiple loci for aluminium resistance. Genome 53:957-966

Rao S, Abdel-Reheem M, Bhella R, McCraken C, Hildebrand D (2008) Characteristics of high α-linolenic acid accumulation in seed oils. Lipids 43:749-755

Rashid KY (2003) Principal diseases of flax. In: Muir AD, Westcott ND (eds) Flax the genus *Linum*. CRC Press, Boca Raton, pp 92-123

Rasmussen LE, Meyer AS (2010) Endogeneous β-D-xylosidase and α-L-arabinofuranosidase activity in flax seed mucilage. Biotechnol Lett 32:1883-1891

Roach MJ, Deyholos MK (2008) Microarray analysis of developing flax hypocotyls identifies novel transcripts correlated with specific stages of phloem fibre differentiation. Ann Bot 102:317-330

Robbins M, Sim S, Yang W, Deynze A, van der Knaap E, Joobeur T, Francis DM (2011) Mapping and linkage disequilibrium analysis with a genome-wide collection of SNPs that detect polymorphism in cultivated tomato. J Exp Bot 62:1831-1845

Roose-Amsaleg C, Cariou-Pham E, Vautrin D, Tavernier R, Solignac M (2006) Polymorphic microsatellite loci in *Linum usitatissimum*. Mol Ecol Notes 6:796-799

Rosenberg NA (2004) Distruct: a program for the graphical display of population structure. Mol Ecol Notes 4:137-138

Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Graner A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc Natl Acad Sci USA 103:18656-18661

Rousset M, Bonnin I, Remoué C, Falque M, Rhoné B, Veyrieras JB, Madur D, Murigneux A, Balfourier F, Le Gouis J, Santoni S, Goldringer I (2011) Deciphering the genetics of flowering time by an association study on candidate genes in bread wheat (*Triticum aestivum* L.). Theor Appl Genet 123:907-926

Rowland GG (1991) An EMS-induced low linolenic acid mutant in McGregor flax (*Linum usitatissimum* L.). Can J Plant Sci 71:393-396

Rowland GG, Bhatty RS (1990) Ethyl methanesulfonate induced fatty acid mutations in flax. J Am Oil Chem Soc 67:213-214

Rowland GG, Hormis YA, Rashid KY (2002) CDC Bethune flax. Can J Plant Sci 82:101-102

Salamini F (2003) Hormones and the green revolution. Science 302:71-72

SAS Institute (2004) SAS Version 9.1. SAS Institute, Cary

Schmegner C, Hoegel J, Vogel W, Assum G (2005) Genetic variability in a genomic region with long-range linkage disequilibrium reveals traces of a bottleneck in the history of the European population. Hum Genet 118:276-286

Schulze T, McMahon F (2002) Genetic association mapping at the crossroads: which test and why? Overview and practical guidelines. Am J Med Genet 114:1-11

Schwarzenbacher H, Dolezal M, Flisikowski K, Seefried F, Wurmser C, Schlötterer C, Fries R (2012) Combining evidence of selection with association analysis increases power to detect regions influencing complex traits in dairy cattle. BMC Genomics 13:48

Sexton TR, Henry RJ, Hardwood CE, Thomas DS, McManus LJ, Raymond C, Henson M, Shepherd M (2011) Pectin methyltransferase genes influence solid wood properties of *Eucalyptus pilularis*. Plant Physiol 158:531-541

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52: 591-611

Shi J, Li R, Qiu D, Jiang C, Long Y, Morgan C, Bancroft I, Zhao J, Meng J (2009) Unraveling the complex trait of crop yield with quantitative trait loci mapping in *Brassica napus*. Genetics 182:851-861

Shimada Y, Shikano T, Merilä J (2011) A high incidence of selection on physiologically important genes in the three-spined stickleback, *Gasterosteus aculeatus*. Mol Biol Evol 28:181-193

Simopoulos AP (2000) Human requirement for N-3 polyunsaturated fatty acids. Poult Sci 79:961–970

Slatkin M (1975) Gene flow and selection in a 2‐locus system. Genetics 81:787–802

Slatkin M (2008) Linkage disequilibrium: Understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477-485

Smith AV, Thomas DJ, Munro HM, Abecasis GR (2005) Sequence features in regions of weak and strong linkage disequilibrium. Genome Res 15:1519-1534

Smooker AM, Wells R, Morgan C, Beaudoin F, Cho K, Fraser F, Bancroft I (2011) The identification and mapping of candidate genes and QTL involved in the fatty acid desaturation pathway in *Brassica napus*. Theor Appl Genet 122:1075-90

Smýkal P, Bačová-Kerteszová N, Kalendar R, Corander J, Schulman AH, Pavelek M (2011) Genetic diversity of cultivated flax (*Linum usitatissimum* L.) germplasm assessed by retrotrasnsposon-based markers. Theor Appl Genet 122:1385-1397

Somerville C, Browse J (1996) Dissecting desaturation: plants prove advantageous. Trends Cell Biol 6:148-153

SØrensen BM, Furukawa-Stoffer TL, Marshall KS, Page EK, Mir Z, Forster RJ, Weselake RJ (2005) Storage lipid accumulation and acyltransferase action in developing flaxseed. Lipids 40:1043-1049

Soto-Cerda BJ, Cloutier S (2012) Association mapping in plant genomes. In: Caliskan M (ed) Genetic diversity in plants. InTech, Rijeka, pp 29–54

Soto-Cerda BJ, Carrasco RJ, Aravena GA, Urbina HA, Navarro CS (2011a) Identifying novel polymorphic microsatellites from cultivated flax (*Linum usitatissimum* L.) following data mining. Plant Mol Biol Rep 29:753-759

Soto-Cerda BJ, Diederichsen A, Ragupathy R, Cloutier S (2013) Genetic characterization of a core collection of flax (*Linum usitatissimum* L.) suitable for association mapping studies and evidence of divergent selection between fiber and linseed types. BMC Plant Biol 13:78

Soto-Cerda BJ, Maureira-Butler I, Muñoz G, Rupayan A, Cloutier S (2012) SSR-based population structure, molecular diversity and linkage disequilibrium analysis of a collection of flax (*Linum usitatissimum* L.) varying for mucilage seed-coat content. Mol Breed 30:875-888

Soto-Cerda BJ, Urbina Saavedra H, Navarro Navarro C, Mora Ortega P (2011b) Characterization of novel genic SSR markers in *Linum usitatissimum* (L.) and their transferability across eleven *Linum* species. Electron J Biotechnol doi: 10.2225/vol14-issue2-fulltext-6

Spielman R, McGinnis R, Ewens W (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506-516

Spielmeyer W, Green AG, Bittisnish D, Mendham N, Lagudah ES (1998) Identification of quantitative trait loci contributing to Fusarium wilt resistance on an AFLP linkage map of flax (*Linum usitatissimum*). Theor Appl Genet 97:633-641

Stanke M, Diekhans M, Baertsch R, Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics 24:637-644

Stapley J, Birkhead T, Burke T, Slate J (2010) Pronounced inter- and intrachromosomal variation in linkage disequilibrium across the zebra finch genome. Genome Res 20:496-502

Stich B, Melchinger A (2010) An introduction to association mapping in plants. CAB Rev Perspect Agric Vet Sci Nutr Nat Resour 5:1-9

Stich B, Piepho HP, Schulz B, Melchinger AE (2008) Multi-traits association mapping in sugar beet (*Beta vulgaris* L.). Theor Appl Genet 117:947-954

Storey JD, Akey JM, Kruglyak L (2005) Multiple locus linkage analysis of genome-wide expression in yeast. PLoS Biol 3:e267

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100:9440-9445

Sundar I (2011) Food security through biodiversity conservation. In: IACSIT (ed) International Conference on Asia Agriculture and Animal, IPCBEE, Singapore, pp 131-138

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739

Tenaillon M, Sawkins M, Long A, Gaut R, Doebley J, Gaut B (2001) Patterns of DNA sequence polymosphism along chromosome 1 of maize (*Zea mays ssp. mays* L.). Proc Natl Acad Sci USA 98:9161-9166

Teo Y, Fry A, Bhattacharya K, Small K, Kwiatkowski D, Clark T (2009) Genome-wide comparisons of variation in linkage disequilibrium. Genome Res 19:1849-1860

The UniProt Consortium (2009) The universal protein resource (UniProt) 2009. Nucleic Acids Res 37:D169-D174

Thornsberry J, Goodman M, Doebley J, Kresovich S, Nielsen D, Buckler E (2001) *Dwarf8* polymorphisms associate with variation in flowering time. Nat Genet 28:286-289

Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet 43:159-162

Touré A, Xueming X (2010) Flaxseed lignans: Source, biosynthesis, metabolism, antioxidant activity, bio-active components, and health benefits. Compr Rev Food Sci F 9:261-269

Tyson H, Fieldes MA, Cheung C, Starobin J (1985) Isozyme relative mobility (Rm) changes related to leaf position; apparently smooth Rm trends and some implications. Biochem Genet 23:641-654

Uysal H, Fu YB, Kurt O, Peterson GW, Diederichsen A, Kusters P (2010) Genetic diversity of cultivated flax (*Linum usitatissimum* L.) and its wild progenitor pale flax

(*Linum bienne* Mill.) as revealed by ISSR markers. Genet Resour Crop Evol 57:1109-1119

van Berloo R (2008) GGT 2.0: Versatile software for visualization and analysis of genetic data. J Hered 99:232-236

van der Merwe R, Labuschagne MT, Herselman L, Hugo A (2013) Stability of seed oil quality traits in high and mid-oleic acid sunflower hybrids. Euphytica 193:157-168

van Zeist W, Bakker-Heeres JAH (1975) Evidence for linseed cultivation before 6000 BC. J Archeol Sci 2:215-219

Vavilov NI (1951) The origin, variation, immunity and breeding of cultivated plants. Chronica Botanica 13:1–366

Venglat P, Xiang D, Qiu S, Stone SL, Tibiche C, Cram D, Alting-Mees M, Nowak J, Cloutier S, Deyholos M, Bekkaoui F, Sharpe A, Wang E, Rowland G, Selvaraj G, Datla R (2011) Gene expression analysis on flax seed development. BMC Plant Biol 11:74

Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y, Doebley J (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proc Natl Acad Sci USA 99:9650-9655

Virk P, Ford-Lloyd B, Jackson M, Pooni H, Clemeno T, Newbury H (1996) Predicting quantitative variation within rice germplasm using molecular markers. Heredity 76:296-304

Vollmann J, Rajcan I (2009) Oil crops breeding and genetics. In: Vollmann J, Rajcan I (eds) Oil crops, handbook of plant breeding 4. Springer, Berlin, pp 1-30

von Kulpa W, Danert S (1962) Zur Systematik von *Linum usitatissimum* L. Kuturpflanze 3:341-388

von Zitzewitz J, Cuesta-Marcos A, Condon F, Castro A, Chao S, Corey A, Filichkin T, Fisk SP, Gutierrez L. Haggard K, Karsai I, Muehlbauer GJ, Smith KP, Veisz O, Hayes PM (2011) The genetics of winter hardiness in barley: perspectives from genome-wide association mapping. Plant Genome 4:76-91

Vrinten P, Hu Z, Munchinsky MA, Rowland G, Qiu X (2005) Two FAD3 desaturase genes control the level of linolenic acid in flax seed. Plant Physiol 139:79-87

VSN International (2011) GenStat for Windows 14th Edition. VSN International, Hemel Hempstead, UK. http://www.GenStat.co.uk

Wang J, McClean P, Lee R, Goos R, Helms T (2008) Association mapping of iron deficiency chlorosis loci in soybean (*Glycine max* L. Merr.) advanced breeding lines. Theor Appl Genet 116:777-787

Wang L, Ge H, Hao C, Dong Y, Zhang X (2012b) Identifying loci influencing 1,000-kernel weight in wheat by microsatellite screening for evidence of selection during breeding. PLoS One 7:e29432

Wang ML, Sukumaran S, Barkley NA, Chen Z, Chen CY, Guo B, Pittman RN, Stalker HT, Holbrook CC, Pederson GA, Yu J (2011) Population structure and marker-trait association analysis of the US peanut (*Arachis hypogaea* L.) mini-core collection. Theor Appl Genet 123:1307-17

Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, Yang L, Hawkins S, Neutelings G, Datla R, Lambert G, Galbraith DW, Grassa CJ, Geraldes A, Cronk QC, Cullis C, Dash PK, Kumar PA, Cloutier S, Sharpe AG, Wong GK, Wang J, Deyholos MK

(2012a) The genome of flax (*Linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. Plant J 72:461-473

Wassom JJ, Wong JC, Martinez E, King JJ, DeBaene J, Hotchkiss JR, Mikkilineni V, Bohn MO, Rocheford TR (2008) QTL associated with maize kernel oil, protein, and starch concentrations; kernel mass; and grain yield in Illinois high oil x B73 backcross-derived lines. Crop Sci 48:243-252

Watterson GA (1978) The homozygosity test of neutrality. Genetics 88:405–417

Weber A, Zhao Q, McMullen M, Doebley J (2009) Using association mapping in teosinte to investigate the function of maize selection-candidate genes. PLoS One 4:e8227

Wen W, Mei H, Feng F, Yu S, Huang Z, Wu J, Chen L, Xu X, Luo L (2009) Population structure and association mapping on chromosome 7 using a diverse panel of Chinese germplasm of rice (*Oryza sativa* L.). Theor Appl Genet 119:459-470

Weselake RJ (2005) Storage lipids. In: Murphy DJ (ed) Plant lipids: biology, utilization and manipulation. Blackwell Publishing, Oxford, pp 162-225

Westcott ND, Muir ND (2003) Chemical studies on the constituents of Linum sp. In: Muir AD, Westcott ND (eds) Flax, the genus Linum. Taylor and Francis, New York, pp 55-73

Western TL, Burn J, Tan WL, Skinner DJ, Martin-McCaffrey L, Moffatt BA, Haughn GW (2001) Isolation and characterization of mutants defective in seed coat mucilage secretory cell development in *Arabidopsis*. Plant Physiol 127:998-1011

Wiesner I, Wiesnerova D, Tejklova E (2001) Effect of anchor and core sequence in microsatellite primers on flax fingerprinting patterns. J Agric Sci 137:37-44

Wiesnerová D, Wiesner I (2004) ISSR-based clustering of cultivated flax germplasm is statistically correlated to thousand seed mass. Mol Biotechnol 26:207-213

Wilson RF (2012) The role of genomics and biotechnology in achieving global food security for high-oleic vegetable oil. J Oleo Sci 61:357-367

Wright S, Gaut B (2005) Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol 22:506-519

Wróbel M, Zebrowski J, Szopa J (2004) Polyhydroxybutyrate synthesis in transgenic flax. J Biotechnol 107:41-54

Würschum T (2012) Mapping QTL for agronomic traits in breeding populations. Theor Appl Genet 125:201-210

Würschum T, Maurer H, Kraf, T, Janssen G, Nilsson C, Reif J (2011) Genome-wide association mapping of agronomic traits in sugar beet. Theor Appl Genet 123:1121-1131

Xiao Y, Cai D, Yang W, Ye W, Younas M, Wu J, Liu K (2012) Genetic structure and linkage disequilibrium pattern of a rapeseed (*Brassica napus* L.) association panel revealed by microsatellites. Theor Appl Genet 125:437-447

Xie D, Han Y, Zeng Y, Chang W, Teng W, Li W (2012) SSR- and SNP-related QTL underlying linolenic acid and other fatty acid contents in soybean seeds across multiple environments. Mol Breed 30:169-179

Xing Y, Zhang Q (2010) Genetic and molecular bases of rice yield. Annu Rev Plant Biol 61:421-42

Xue W, Xing Y, Weng X, Zhao Y, Tang W, Wang L, Zhou H, Yu S, Xu C, Li X, Zhang Q (2008) Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. Nat Genet 143:1-7

Yan J, Shan T, Warburton M, Buckler E, McMullen M, Crouch J (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PloS One 4:e8451

Yan W, Tinker NA (2005) A biplot approach for investigating QTL-by-environment patterns. Mol Breed 15:31-43

Yang X, Yan J, Shah T, Warburton ML, Li Q, Li L, Gao Y, Chai Y, Fu Z, Zhou Y, Xu S, Bai G, Meng Y, Zheng Y, Li J (2010) Genetic analysis and characterization of a new maize association mapping panel for quantitative trait loci dissection. Theor Appl Genet 121:417-31

Yeh FC, Yang RC, Boyle TBJ, Ye ZH, Mao JX (1997) POPGENE, the user-friendly shareware for population genetic analysis. Molecular biology and biotechnology centre: University of Alberta Press, Edmonton

Yin T, DiFazio S, Gunter L, Jawdy S, Boerjan W, Tuskan G (2004) Genetic and physical mapping of *Melampsora* rust resistance genes in *Populus* and characterization of linkage disequilibrium and flanking genomic sequence. New Phytologist 164:95-105

You FM, Duguid SD, Thambugala D, Cloutier S (2013) Statistical analysis and field evaluation of the type 2 modified augmented design in phenotyping of flax germplasm in multiple environments. Aust J Crop Sci 7:1789-1800

Yu J, Buckler E (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17:155-160

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539-551

Yu L, Lorenz A, Rutkoski J, Singh R, Bhavani S, Huerta-Espino J, Sorrells ME (2011) Association mapping and gene-gene interaction for stem rust resistance in CIMMYT spring wheat germplasm. Theor Appl Genet 123:1257-1268

Yu J, Pressoir G, Briggs W, Vroh Bi I, Yamasaki M, Doebley J, McMullen M, Gaut B, Nielsen D, Holland J, Kresovich S, Buckler E (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203-208

Yunusbayev B, Metspalu M, Järve M, Kutuev I, Rootsi S, Metspalu E, Behar DM, Varendi K, Sahakyan H, Khusainova R, Yepiskoposyan L, Khusnutdinova EK, Underhill PA, Kivisild T, Villems R (2011) The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Mol Biol Evol 29:359-365

Zhang D, Bai G, Zhu C, Yu J, Carver B (2010a) Genetic diversity, population structure, and linkage disequilibrium in U.S. elite winter wheat. Plant Genome 3:117-127

Zhang D, Hao C, Wang L, Zhang X (2012) Identifying loci influencing grain number by microsatellite screening in bread wheat (*Triticum aestivum* L.). Planta 236:1507-1517

Zhang LY, Liu DC, Guo XL, Yang WL, Sun JZ, Wang DW, Zhang A (2010b) Genomic distribution of quantitative trait loci for yield and yield-related traits in common wheat. J Integr Plant Biol 52:996-1007

Zhao J, Becker HC, Zhang D, Zhang X, Ecke W (2005) Oil content in a European × Chinese rapeseed population: QTL with additive and epistatic effects and their genotype–environment interactions. Crop Sci 45: 51-59

Zhao J, Paulo MJ, Jamar D, Lou P, van Eeuwijk F, Bonnema G, Vreugdenhil D, Koornneef M (2007a) Association mapping of leaf traits, flowering time, and phytate content in *Brassica rapa*. Genome 50:963-973

Zhao K, Aranzana M, Kim S (2007b) An Arabidopsis example of association mapping in structured samples. PLoS Genet 3:e4

Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat Commun 2:467

Zhu C, Gore M, Buckler E, Yu J (2008) Status and prospects of association mapping in plants. The Plant Genome 1:5-20

Zhu Q, Zheng X, Luo J, Gaut B, Ge S (2007) Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. Mol Biol Evol 24:875-888

Zobel RW, Wright MG, Gauch HG (1988) Statistical analysis of yield trial. Agron J 80:388-393

Zou J, Jiang C, Cao Z, Li R, Long Y, Chen S, Meng J (2010) Association mapping of seed oil content in *Brassica napus* and comparison with quantitative trait loci identified from linkage mapping. Genome 53:908-916

APPENDICES

Appendix I Association mapping studies in plants.

Species	Germplasm	Trait	Marker system	Reference
Arabidopsis	Diverse accessions	Flowering time/pathogen resistance	Sequences	Aranzana et al. (2005)
	Diverse accessions	Multiple traits	SSRs/SNPs	Ersoz et al. (2007)
	Natural accessions	Flowering time	SNPs	Brachi et al. (2010)
	Diverse accessions	Climate-sensitive QTL	SNPs	Li et al. (2010)
	Landraces	Downy mildew	SNPs	Nemri et al. (2010)
Maize	Inbred lines	Aluminum tolerance	SNPs	Krill et al. (2010)
	Inbred lines	Drought tolerance	SNPs	Lu et al. (2010)
	Inbred lines	Northern leaf blight	SNPs	Poland et al. (2011)
	Inbred lines	Southern leaf blight	SNPs	Kump et al. (2011)
	Inbred lines	Leaf architecture	SNPs	Tian et al. (2011)
Teosinte	Landraces	Domestication-related genes	SNPs	Weber et al. (2009)
Wheat	Cultivars	Kernel size, milling quality	SSRs	Breseghello and Sorrells (2006)
	Diverse accessions	Aluminum resistance	DArT	Raman et al. (2010)
	Breeding lines	Stem rust resistance	DArT	Yu et al. (2011)
	Diverse accessions	Flowering time	SNPs	Rousset et al. (2011)
Barley	Inbred lines	Growth habit	SNPs	Rostoks et al. (2006)
	Cultivars	Anthocyanin pigmentation	SNPs	Cockram et al. (2010)
	Breeding lines	Winter hardiness	SNPs	Von Zitzewitz et al. (2011)
	Spring cultivars	Multiple traits	SNPs	Pasam et al. (2012)

Species	Germplasm	Trait	Marker system	Reference
Oat	Diverse cultivars	Agronomic and kernel quality traits	AFLPs	Achleitner et al. (2008)
Rice	Landraces	Heading date, plant height and panicle length	SSRs	Wen et al. (2009)
	Landraces	Multiple agronomic traits	SNPs	Huang et al. (2010)
	Cultivars, subspecies	Flowering time, grain yield	SNPs	Huang et al. (2012)
Canola	Diverse accessions	Leaf traits, flowering time and phytate content	AFLPs	Zhao et al. (2007)
	Diverse accessions	Oil content	SSRs	Zou et al. (2010)
Soybean	Breeding lines	Iron deficiency chlorosis	SSRs	Wang et al. (2008)
Cotton	Diverse cultivars	Fiber quality	SSRs	Abdurakhmonov et al. (2009)
Peanut	Diverse accessions	Seed quality traits	SSRs-SNPs	Wang et al. (2011)
Sugar beet	Inbred lines	Sugar content and yield	SSRs	Stich et al. (2008)
Sugar beet	Inbred lines	Multiple traits	SNPs	Würschum et al. (2011)
Alfalfa	Cultivars	Biomass yield and stem composition	SSRs	Li et al. (2011a)
Sunflower	Inbred lines	Flowering time	SNPs	Cadic et al. (2013)

SNPs: Single Nucleotide Polymorphisms; SSRs: Simple Sequence Repeats; DArT: Diversity Arrays Technology; AFLPs: Amplified Fragment Length Polymorphisms

Appendix II List of software used in LD and AM analyses.

Software	Focus	Description	Website
STRUCTURE 2.3	Population structure	Compute a MCMC Bayesian analysis to estimate the proportion of the genome of an individual originating from the different inferred populations	http://pritch.bsd.uchicago.edu/software.html
BAPS 5.0	Population structure	Compute Bayesian analysis to estimate the proportion of the genome of an individual and assign individuals to genetic clusters by either considering them as immigrants or as descendants from immigrants	http://web.abo.fi/fak/mnf/mate/jc/software/baps.html
mStruct	Population structure	Detection of population structure in the presence of admixing and mutations from multi-locus genotype data. It is a mixed membership model (also referred to as an admixture model) which incorporates a mutation process on the observed genetic markers	http://www.cs.cmu.edu/~suyash/mstruct.html
LDheatmap	LD	R environment software for LD estimation (r^2) displayed as heatmap plot using SNPs	http://www.jstatsoft.org/v16/c03
LDhat 2.1	Recombination rates and LD	R environment software for LD estimation and identification of hotspot using a Bayesian reversible jump MCMC a scheme for SNPs	www.stats.ox.ac.uk/mcvean/ldhat.html

Software	Focus	Description	Website
Arlequin 3.5	Genetic analysis and LD	Hierarchical analysis of genetic structure (AMOVA), LD for D' and r^2. Incorporates a R function to parse XML output files to produce publication quality graphics	http://cmpg.unibe.ch/software/arlequin35/
Haploview 4.2	Haplotype analysis and LD	LD and haplotype block analysis, haplotype population frequency estimation, single SNP and haplotype association tests, permutation testing for association significance	www.broad.mit.edu/mpg/haploview/
GGT 2.0	Genetic analysis, LD and AM	Compute genetic distance based on Jaccard similarity, dendrograms are displayed using Neighbor-Joining algorithm. Displays LD heatmaps and LD scatter plots for D' and r^2 and performs simple AM analysis	http://www.plantbreeding.wur.nl/UK/software_ggt.html
SVS 7	Stratification, LD and AM	Estimate stratification, LD, haplotypes blocks and multiple AM approaches for up to 1.8 million SNPs and 10,000 samples	www.goldenhelix.com
TASSEL	Stratification, LD and AM	SSR markers, GLM and MLM methods	http://www.maizegenetics.net
GenStat	Stratification, LD and AM	SSR markers, GLM and MLM-PCA methods	http://www.vsni.co.uk/

Software	Focus	Description	Website
GenAMap	Stratification, LD and structured AM	SNPs, tree of functional branches, multiple visualization tools	http://cogito-b.ml.cmu.edu/genamap
PLINK	Stratification, LD and structured AM	SNPs, multiple AM approaches, IBD and IBS analyses	http://pngu.mgh.harvard.edu/purcell/plink/

Appendix III Core collection data including accession number, accession name, type, improvement status and origin.

Sample	Accession number	Name	Type	Improvement status	Origin
1	CN18973	AC WATSON	Oil	Cultivar	CAN
2	CN18979	FLANDERS	Oil	Cultivar	CAN
3	CN18980	SOMME	Oil	Cultivar	CAN
4	CN18981	CDC VALOUR	Oil	Cultivar	CAN
5	CN18982	EVELIN	Fiber	Cultivar	FRA
6	CN18983	LAURA	Fiber	Cultivar	NLD
7	CN18986	HERMES	Fiber	Cultivar	FRA
8	CN18987	VIKING	Fiber	Cultivar	NLD
9	CN18988	ARIANE	Fiber	Cultivar	FRA
10	CN18989	ATALANTE	Oil	Cultivar	FRA
11	CN18991	NIKE	Fiber	Cultivar	POL
12	CN18993	LINDA	Oil	Cultivar	NLD
13	CN18994	VERNE	Oil	Cultivar	USA
14	CN18997	RAISA	Fiber	Cultivar	NLD
15	CN18998	ESCALINA	Fiber	Cultivar	NLD
16	CN19001	MARINA	Fiber	Cultivar	NLD
17	CN19003	AC MCDUFF	Oil	Cultivar	CAN
18	CN19004	AC EMERSON	Oil	Cultivar	CAN
19	CN19005	AC LINORA	Oil	Cultivar	CAN
20	CN19007	LIN-1724	Oil	Breeding material	ETH
21	CN19017	CDC NORMANDY	Oil	Cultivar	CAN
22	CN19157	OTTAWA 829-C	Oil	Cultivar	CAN
23	CN19158	OTTAWA 770B	Oil	Cultivar	CAN

Sample	Accession number	Name	Type	Improvement status	Origin
24	CN19159	DIADEM	Oil	Cultivar	CAN
25	CN19160	BOLLEY GOLDEN	Oil	Cultivar	USA
26	CN30860	Kirovogradskij 71	Oil	Cultivar	UKR
27	CN30861	Kubanskij	Oil	Cultivar	UNK
28	CN32542	VNIIL-17	Fiber	Cultivar	RUS
29	CN32546	Korostenskij 3	Fiber	Cultivar	UKR
30	CN33385	LINOTT	Oil	Cultivar	CAN
31	CN33386	NORALTA	Oil	Cultivar	CAN
32	CN33388	REDWOOD 65	Oil	Cultivar	CAN
33	CN33389	ROCKET	Oil	Cultivar	CAN
34	CN33390	NATASJA	Fiber	Cultivar	NLD
35	CN33393	Domtar Selection	Fiber	Cultivar	UNK
36	CN33397	DUFFERIN	Oil	Cultivar	CAN
37	CN33399	BISON	Oil	Cultivar	USA
38	CN33400	NORSTAR	Oil	Cultivar	USA
39	CN33992	CULBERT	Oil	Cultivar	USA
40	CN35791	TVERCA	Fiber	Cultivar	RUS
41	CN37286	MCGREGOR	Oil	Cultivar	CAN
42	CN40081	NATASJA	Fiber	Cultivar	NLD
43	CN52732	NORLIN	Oil	Cultivar	CAN
44	CN96845	CIli-642	Oil	Cultivar	RUS
45	CN96846	CIli-643	Oil	Cultivar	RUS
46	CN96911	CIli-1407	Oil	Cultivar	TUR

Sample	Accession number	Name	Type	Improvement status	Origin
47	CN96958	CIli-1455	Oil	Landrace	TUR
48	CN96962	CIli-1458	Oil	Cultivar	TUR
49	CN96974	CIli-1470	Oil	Landrace	IND
50	CN96988	CIli-1499	Oil	Cultivar	ETH
51	CN96991	CIli-1502	Oil	Cultivar	ETH
52	CN96992	CIli-1503	Oil	Cultivar	ETH
53	CN97004	CIli-1519	Oil	Cultivar	ETH
54	CN97050	CIli-1924	Oil	Cultivar	IRN
55	CN97056	CIli-1930	Oil	Cultivar	PAK
56	CN97064	CIli-1938	Oil	Cultivar	PAK
57	CN97072	CIli-1946	Oil	Landrace	PAK
58	CN97083	CIli-1957	Oil	Landrace	PAK
59	CN97092	CIli-1991	Oil	Cultivar	PAK
60	CN97096	CIli-1995	Oil	Cultivar	PAK
61	CN97103	CIli-2002	Oil	Cultivar	PAK
62	CN97129	CIli-2028	Oil	Landrace	IRN
63	CN97129B	CIli-2028B	Oil	Landrace	IRN
64	CN97139	CIli-2038	Oil	Cultivar	IRN
65	CN97147	CIli-2046	Oil	Cultivar	TUR
66	CN97153	CIli-2052	Oil	Cultivar	TUR
67	CN97176	HORAL	Oil	Cultivar	CZE
68	CN97180	Sorth Behbehan	Fiber	Cultivar	IRN
69	CN97214	CIli-2295	Oil	Cultivar	ARG

Sample	Accession number	Name	Type	Improvement status	Origin
70	CN97238	No. 1048	Oil	Cultivar	HUN
71	CN97287	Lina Deta	Oil	Cultivar	HUN
72	CN97300	RAJA	Oil	Cultivar	HUN
73	CN97306	N.P. (R.R.) 9	Oil	Cultivar	IND
74	CN97307	N.P. (R.R.) 37	Oil	Cultivar	IND
75	CN97308	N.P. (R.R.) 38	Oil	Cultivar	IND
76	CN97312	T. 126	Oil	Cultivar	IND
77	CN97321	CIli-2528	Oil	Cultivar	ROM
78	CN97334	MOCORETA	Oil	Cultivar	ARG
79	CN97341	H723 F3-6-3-3-4-2-2	Unknown	Cultivar	ARG
80	CN97350	de metcha 1-3-3 Vilm	Oil	Cultivar	FRA
81	CN97351	de metcha 1-3-6 Vilm	Fiber	Cultivar	FRA
82	CN97366	Texas S. 4-6 Walsh x New Golden	Oil	Cultivar	USA
83	CN97377	Reserve (N. Dak. Res. 155)	Oil	Cultivar	USA
84	CN97392	NOVELTY	Oil	Cultivar	CAN
85	CN97393	Sel. C.I. 21-2 Jalaun	Oil	Cultivar	USA
86	CN97396	Res. x Hoshangabad (C.I. 19 x C.I. 140)	Oil	Cultivar	USA
87	CN97397	Sel. C.I. 19-47 Pale Blue	Unknown	Cultivar	USA
88	CN97402	No. Dak. No. 40,013	Unknown	Breeding material	USA
89	CN97403	LINOTA	Oil	Cultivar	USA
90	CN97404	Buda Sel.	Oil	Breeding material	USA
91	CN97404B	Buda Sel.B	Oil	Breeding material	USA
92	CN97406	No.Dak. Res. No. 52	Unknown	Breeding material	USA

Sample	Accession number	Name	Type	Improvement status	Origin
93	CN97407	Rio (Long 79)	Oil	Cultivar	USA
94	CN97424	Tammes #3 White Involute	Fiber	Cultivar	NLD
95	CN97430	N.D. Nur. No. 1740 (G.36 a/21)	Oil	Breeding material	DEU
96	CN97430B	N.D. Nur. No. 1740 (G.36 a/21)B	Oil	Breeding material	DEU
97	CN97444	N.D. Resistant 714	Oil	Cultivar	USA
98	CN97452	Sel. of Minn. 281	Unknown	Cultivar	USA
99	CN97453	Pale Blue Sel. from N.D.R. 52	Unknown	Cultivar	USA
100	CN97458	CIli-469	Oil	Cultivar	NLD
101	CN97463	Sel. of N.D.R. 114	Oil	Cultivar	USA
102	CN97470	Sagino	Oil	Cultivar	JPN
103	CN97475	Common White	Oil	Cultivar	RUS
104	CN97483	CIli-522	Unknown	Landrace	RUS
105	CN97484	CIli-523	Oil	Cultivar	RUS
106	CN97487	CIli-526	Oil	Cultivar	RUS
107	CN97489	CIli-531	Oil	Cultivar	RUS
108	CN97503	CIli-556	Fiber	Landrace	RUS
109	CN97520	CIli-576	Oil	Cultivar	RUS
110	CN97529	CIli-589	Oil	Cultivar	RUS
111	CN97530	CIli-590	Fiber	Landrace	RUS
112	CN97531	CIli-593	Fiber	Landrace	RUS
113	CN97533	CIli-595	Fiber	Landrace	RUS
114	CN97571	Cyprus	Oil	Cultivar	CAN
115	CN97584	Minn. Sel. Winona x 770B F5	Unknown	Breeding material	USA

Sample	Accession number	Name	Type	Improvement status	Origin
116	CN97584B	Minn. Sel. Winona x 770B F6-B	Unknown	Breeding material	USA
117	CN97586	Long 66 (non-ciliate)	Oil	Cultivar	USA
118	CN97587	Capa (Argentine)	Oil	Cultivar	USA
119	CN97604	CIli-758	Oil	Cultivar	RUS
120	CN97605	CIli-759	Oil	Landrace	RUS
121	CN97610	Tammes Type 2	Fiber	Cultivar	NLD
122	CN97613	Tammes Type 5	Oil	Cultivar	NLD
123	CN97616	Tammes Type 12	Fiber	Cultivar	NLD
124	CN97633	Royal	Oil	Cultivar	CAN
125	CN97639	CIli-835	Oil	Cultivar	USA
126	CN97639	CIli-835B	Oil	Cultivar	USA
127	CN97642	Renew	Oil	Cultivar	USA
128	CN97649	CIli-847	Oil	Cultivar	USA
129	CN97665	CIli-854	Fiber	Breeding material	USA
130	CN97670	No. 5242 - 1937	Unknown	Breeding material	USA
131	CN97671	J.W.S.	Oil	Cultivar	CAN
132	CN97679	CIli-897	Oil	Breeding material	USA
133	CN97679B	CIli-897B	Oil	Breeding material	USA
134	CN97689	CIli-908	Oil	Cultivar	USA
135	CN97718	CIli-946	Oil	Cultivar	USA
136	CN97728	CIli-956	Oil	Cultivar	USA
137	CN97740	Redson	Oil	Cultivar	USA
138	CN97749	Crystal	Oil	Cultivar	USA

Sample	Accession number	Name	Type	Improvement status	Origin
139	CN97768	Mourisco, E730	Unknown	Cultivar	PRT
140	CN97871	Atlas (fiber)	Fiber	Cultivar	SWE
141	CN97873	Redwood (C.I. 980 x Redson)	Oil	Cultivar	USA
142	CN97881	Biwing x C.I. 980 (II-40-35)	Oil	Cultivar	USA
143	CN97886	Lusatia	Oil	Cultivar	DEU
144	CN97890	Maritime	Oil	Cultivar	USA
145	CN97907	Victory B	Oil	Cultivar	USA
146	CN97921	CIli-1185	Oil	Cultivar	USA
147	CN97953	10382/46	Oil	Cultivar	ARG
148	CN97958	10387/46	Oil	Cultivar	ARG
149	CN97961	10390/46	Oil	Cultivar	ARG
150	CN97967	10397/46	Oil	Cultivar	ARG
151	CN97980	10410/46	Oil	Cultivar	ARG
152	CN98007	10442/46	Oil	Cultivar	ARG
153	CN98012	10447/46	Oil	Cultivar	ARG
154	CN98014	10451/46	Oil	Cultivar	ARG
155	CN98027	10469/46	Oil	Cultivar	ARG
156	CN98037	10479/46	Oil	Breeding material	ARG
157	CN98037B	10479/46B	Oil	Breeding material	ARG
158	CN98039	10481/46	Oil	Cultivar	ARG
159	CN98056	Hollandia	Oil	Cultivar	NLD
160	CN98056B	Hollandia-B	Oil	Cultivar	NLD
161	CN98057	CIli-1474	Oil	Cultivar	IND

Sample	Accession number	Name	Type	Improvement status	Origin
162	CN98072	Unryu	Fiber	Cultivar	JPN
163	CN98100	Uruguay 36/48	Oil	Cultivar	URY
164	CN98109	CIli-1562	Oil	Cultivar	IND
165	CN98135	CIli-1596	Oil	Cultivar	IND
166	CN98150	Z 11637	Fiber	Unknown	NLD
167	CN98157	R.R. 38	Oil	Cultivar	IND
168	CN98165	1546-S	Oil	Cultivar	IRN
169	CN98176	1224-S	Oil	Cultivar	AFG
170	CN98192	CIli-1653	Oil	Cultivar	IRL
171	CN98193	L.G. 0189B	Unknown	Cultivar	MAR
172	CN98231	CIli-1749	Oil	Cultivar	USA
173	CN98237	CIli-1827	Oil	Cultivar	PAK
174	CN98239	CIli-1829	Oil	Cultivar	PAK
175	CN98240	CIli-1830	Oil	Landrace	IND
176	CN98240B	CIli-1830B	Oil	Landrace	IND
177	CN98242	CIli-1832	Oil	Landrace	IND
178	CN98250	CIli-1840	Oil	Cultivar	IND
179	CN98254	Basin	Oil	Cultivar	IND
180	CN98263	Chaurra Olajlen	Oil	Cultivar	HUN
181	CN98263	Chaurra Olajlen-B	Oil	Cultivar	HUN
182	CN98275	N 39/a La Plata	Oil	Cultivar	HUN
183	CN98276	N 39/b La Plata	Oil	Cultivar	HUN
184	CN98278	Karnobat 4	Oil	Cultivar	HUN

Sample	Accession number	Name	Type	Improvement status	Origin
185	CN98279	Karnobat 5	Oil	Cultivar	ARG
186	CN98286	Mapun	Fiber	Cultivar	HUN
187	CN98303	Torok 11	Fiber	Cultivar	HUN
188	CN98363	N.P. 30	Oil	Cultivar	IND
189	CN98364	N.P. 31	Oil	Cultivar	IND
190	CN98370	N.P. 37	Oil	Cultivar	IND
191	CN98397	N.P. 65	Oil	Cultivar	IND
192	CN98398	N.P. 66	Oil	Cultivar	IND
193	CN98415	N.P. 86	Oil	Cultivar	IND
194	CN98440	N.P. 109	Oil	Cultivar	IND
195	CN98467	N.P. (RR.) 405	Oil	Cultivar	IND
196	CN98468	N.P. (RR.) 407	Oil	Cultivar	IND
197	CN98475	Flachskopf	Oil	Cultivar	DEU
198	CN98505	Varoneshski 1308	Oil	Cultivar	RUS
199	CN98535	Texas S. 32-1 Viking x Norsk	Oil	Cultivar	USA
200	CN98541	Texas S. 32-1 Viking x Norsk	Oil	Cultivar	USA
201	CN98542	Amalla' H.D. Long Sel.	Oil	Cultivar	USA
202	CN98566	Rwd x Mar Minn. 61-2151	Oil	Breeding material	USA
203	CN98566B	Rwd x Mar Minn. 61-2151-B	Oil	Breeding material	USA
204	CN98566		Oil	Breeding material	USA
205	CN98569	CIli-2473	Oil	Unknown	IND
206	CN98610	Brawley R0001 (yel.sd. sel.)	Unknown	Cultivar	USA
207	CN98613	Br.B502 (Imp. x Punj. 473) 1008-2	Oil	Cultivar	USA

Sample	Accession number	Name	Type	Improvement status	Origin
208	CN98634	Toba	Oil	Cultivar	ARG
209	CN98639	W5618RO-41	Oil	Cultivar	USA
210	CN98644	W5623RO-24	Unknown	Breeding material	USA
211	CN98683	Mapum M.A.	Oil	Cultivar	CZE
212	CN98689	Primus	Unknown	Cultivar	CZE
213	CN98704	Wicking Hegenan	Fiber	Cultivar	CZE
214	CN98708	Vitagold	Fiber	Cultivar	FRA
215	CN98710	Erythree	Fiber	Landrace	FRA
216	CN98712	Safi 1.4-2-1	Oil	Cultivar	FRA
217	CN98733	Bulgare a h	Oil	Breeding material	POL
218	CN98734	Karnobat 9	Oil	Cultivar	FRA
219	CN98741	Karnobat 1591 1.9	Oil	Breeding material	FRA
220	CN98742	Comun de Diaz	Unknown	Landrace	FRA
221	CN98752	Lina grosses graines Vilmorin No1	Oil	Cultivar	FRA
222	CN98753	Lina de Safi Vilmorin No2	Oil	Cultivar	FRA
223	CN98767	LG 0196	Oil	Cultivar	FRA
224	CN98773	Safi 1.1-2-5	Oil	Cultivar	FRA
225	CN98794	Lino de Cabiro	Oil	Cultivar	FRA
226	CN98806	CIli-2761	Oil	Breeding material	FRA
227	CN98807	CIli-2762	Oil	Cultivar	FRA
228	CN98812	Bison LN (67-I-46)	Oil	Cultivar	USA
229	CN98821	Foster ND14a (1605 x Minerva)	Oil	Cultivar	USA
230	CN98826	Common	Fiber	Cultivar	EGY

Sample	Accession number	Name	Type	Improvement status	Origin
231	CN98829	Dolgunetz	Fiber	Cultivar	USA
232	CN98854	CIli-1573	Oil	Cultivar	HUN
233	CN98869	Field N. 17	Oil	Cultivar	TUR
234	CN98903	CIli-1753	Fiber	Breeding material	USA
235	CN98923	CIli-1774	Fiber	Breeding material	USA
236	CN98926	CIli-1777	Fiber	Breeding material	USA
237	CN98934	Wada Fiber	Fiber	Cultivar	USA
238	CN98946	Talmune Fiber	Fiber	Cultivar	USA
239	CN98954	Cascade Fiber	Fiber	Cultivar	USA
240	CN98961	CIli-1846	Oil	Cultivar	IND
241	CN98969	N.P. 15	Oil	Cultivar	IND
242	CN98973	N.P. 117	Oil	Cultivar	IND
243	CN98974	N.P. 118	Oil	Cultivar	IND
244	CN98982	N.P. (RR.) 272	Oil	Cultivar	IND
245	CN98984	Bonnydoon-9 (H39-9)	Oil	Cultivar	AUS
246	CN100547	Redwing	Oil	Cultivar	UNK
247	CN100629	CIli-2971	Oil	Cultivar	PAK
248	CN100674	CIli-3026	Oil	Cultivated material	ROM
249	CN100678	CIli-3030	Unknown	Cultivated material	ROM
250	CN100770	CIli-3250	Oil	Breeding material	USA
251	CN100785	VERNE 93 SDT8914	Oil	Breeding material	USA
252	CN100790	Ghari 3	Unknown	Cultivated material	PAK
253	CN100795	Tammes Pale Blue	Fiber	Cultivar	NLD

Sample	Accession number	Name	Type	Improvement status	Origin
254	CN100797	SP 2271	Unknown	Breeding material	NZL
255	CN100797B	SP 2271-B	Unknown	Breeding material	NZL
256	CN100799	N.P. 84	Unknown	Cultivated material	IND
257	CN100805	Floribus Roseis	Oil	Cultivar	CZE
258	CN100807	LIN-1062	Oil	Unknown	AFG
259	CN100827	Safedak	Oil	Cultivar	SUN
260	CN100828	Winterlein	Unknown	Unknown	TUR
261	CN100837	LIN-1193	Oil	Unknown	TUR
262	CN100838	LIN-706	Oil	Unknown	CYP
263	CN100841	LIN-627	Oil	Unknown	UNK
264	CN100848	Ottawa 2152	Fiber	Cultivar	CAN
265	CN100851	Sumpersky Fa 13 Jenny	Oil	Cultivar	CZE
266	CN100852	Grandal	Unknown	Cultivar	PRT
267	CN100863	LIN-771	Oil	Breeding material	FRA
268	CN100864	Bjelo Katjacs	Fiber	Cultivar	HUN
269	CN100881	Deutscher Ollein	Oil	Cultivar	DEU
270	CN100883	Beta 201	Oil	Cultivar	HUN
271	CN100884	g. 12 Ruzokvety	Oil	Cultivar	CSK
272	CN100885	aus Lathyrus	Unknown	Unknown	GRC
273	CN100895	Karbin (landrace)	Unknown	Landrace	ETH
274	CN100910	Grandal (landrace)	Oil	Landrace	PRT
275	CN100928	Ocean	Oil	Cultivar	FRA
276	CN100929	Belinka	Fiber	Cultivar	NLD

Sample	Accession number	Name	Type	Improvement status	Origin
277	CN100939	VNIIL-7939	Oil	Cultivar	RUS
278	CN100952	VIR-1270	Fiber	Unknown	AFG
279	CN101016	Zheltosemyannyi	Unknown	Cultivar	CHN
280	CN101026	6 V27-2 (pop.varieta)	Oil	Breeding material	MAR
281	CN101038	Nika	Fiber	Cultivar	BLR
282	CN101039	VNIIL-4767	Fiber	Breeding material	RUS
283	CN101052	L-93-2	Fiber	Breeding material	CHN
284	CN101053	L-8709-5-10	Fiber	Breeding material	CHN
285	CN101055	L-140-16	Fiber	Breeding material	RUS
286	CN101094	Torzhokskij 4	Fiber	Cultivar	RUS
287	CN101096	Novotorzhskij	Fiber	Cultivar	RUS
288	CN101099	Aleksim	Fiber	Cultivar	RUS
289	CN101114	VNIIL-5320	Fiber	Breeding material	RUS
290	CN101115	VNIIL-5321	Fiber	Breeding material	RUS
291	CN101116	VNIIL-5316	Fiber	Breeding material	RUS
292	CN101118	VNIIL-5316	Fiber	Breeding material	LTU
293	CN101119	VNIIL-3177	Fiber	Breeding material	RUS
294	CN101127	VNIIL-5317	Fiber	Breeding material	RUS
295	CN101132	VNIIL-5623	Oil	Breeding material	RUS
296	CN101136	Verchnevolzhkij	Fiber	Cultivar	RUS
297	CN101137	VNIIL-5613	Oil	Breeding material	RUS
298	CN101154	Belochka	Fiber	Cultivar	RUS
299	CN101208	VNIIL-5545	Oil	Cultivar	IND

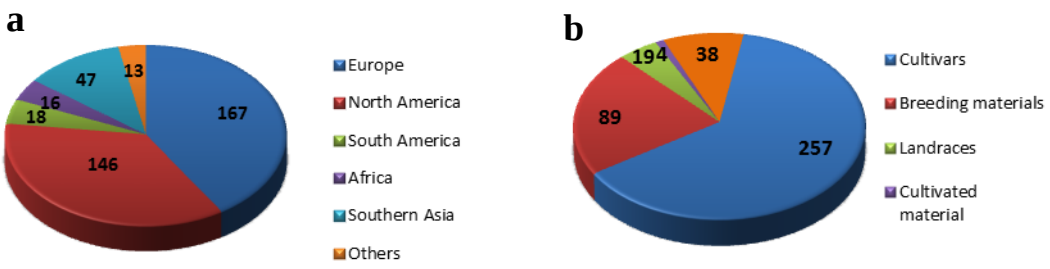
Sample	Accession number	Name	Type	Improvement status	Origin
300	CN101230	VNIIL-5520	Fiber	Breeding material	CHN
301	CN101237	Artemida	Oil	Cultivar	LTU
302	CN101240	VNIIL-6182	Oil	Breeding material	LTU
303	CN101241	VNIIL-5325	Oil	Breeding material	RUS
304	CN101265	Amason	Oil	Cultivar	GBR
305	CN101279	VNIIL-5606	Oil	Breeding material	RUS
306	CN101286	Dakota Line 8	Oil	Breeding material	USA
307	CN101289	VNIIL-5680	Oil	Breeding material	RUS
308	CN101296	L. 270-68	Oil	Breeding material	RUS
309	CN101298	L. 541-02	Oil	Breeding material	RUS
310	CN101299	L. 00-207	Oil	Breeding material	RUS
311	CN101301	L. 1200-4-3	Oil	Breeding material	RUS
312	CN101307	LM-95	Oil	Breeding material	RUS
313	CN101308	VNIIL-180	Oil	Unknown	IND
314	CN101310	VNIIL-571	Oil	Unknown	IND
315	CN101325	VNIIL-1104	Oil	Unknown	GRC
316	CN101327	VNIIL-6148	Unknown	Unknown	ESP
317	CN101329	VNIIL-519	Oil	Unknown	EGY
318	CN101331	VNIIL-918	Oil	Unknown	TUR
319	CN101332	VNIIL-1046	Oil	Unknown	TUR
320	CN101338	VNIIL-655	Oil	Unknown	AFG
321	CN101348	VNIIL-742	Fiber	Unknown	RUS
322	CN101364	VNIIL-776	Fiber	Unknown	RUS

Sample	Accession number	Name	Type	Improvement status	Origin
323	CN101366	VNIIL-725	Oil	Unknown	GEO
324	CN101367	VNIIL-2785	Oil	Unknown	GEO
325	CN101373	VNIIL-868	Oil	Unknown	ARM
326	CN101375	VNIIL-3531	Oil	Unknown	RUS
327	CN101378	VNIIL-409	Fiber	Unknown	UKR
328	CN101379	VNIIL-492	Fiber	Unknown	UKR
329	CN101382	Keteni Tekirdak Hagrobobu	Fiber	Unknown	TUR
330	CN101385	TR 35141	Fiber	Unknown	TUR
331	CN101386	TR 42713	Fiber	Unknown	TUR
332	CN101392	Tajga	Fiber	Cultivar	FRA
333	CN101394	Line 548-01	Fiber	Unknown	RUS
334	CN101395	Line 629-01	Fiber	Unknown	RUS
335	CN101396	Line 657-01	Fiber	Unknown	RUS
336	CN101397	Pskovski 2976	Fiber	Unknown	UKR
337	CN101401	G 2063-5-10	Fiber	Unknown	RUS
338	CN101402	VNIIL-5631	Fiber	Unknown	RUS
339	CN101403	L-500004-2-84	Fiber	Unknown	ROM
340	CN101404	L-60016-3-87	Fiber	Unknown	ROM
341	CN101405	Mures	Fiber	Unknown	ROM
342	CN101406	L-41	Fiber	Unknown	RUS
343	CN101407	Concurent	Fiber	Unknown	NLD
344	CN101413	Vimy	Oil	Cultivar	CAN
345	CN101416	China 1	Fiber	Breeding material	CHN

Sample	Accession number	Name	Type	Improvement status	Origin
346	CN101417	China 2	Fiber	Breeding material	CHN
347	CN101419	China 4	Fiber	Breeding material	CHN
348	CN101421	China 5	Fiber	Breeding material	CHN
349	CN101448	Sel CIli -332 (C5)	Oil	Breeding material	CAN
350	CN101451	Sel CIli -400 (C5)	Unknown	Breeding material	CAN
351	CN101454	Sel. of CIli-684 (C4)	Oil	Breeding material	CAN
352	CN101461	Sel. of CIli-1020 (C4)	Oil	Breeding material	CAN
353	CN101463	Sel. of CIli-1220 (C4)	Oil	Breeding material	CAN
354	CN101466	Sel. of CIli-1472 (C4)	Oil	Breeding material	CAN
355	CN101469	Sel. of CIli-1484 (C4)	Oil	Breeding material	CAN
356	CN101471	Sel CIli -1490 (C4)	Oil	Breeding material	CAN
357	CN101472	Sel. of CIli-1493 (C4)	Oil	Breeding material	CAN
358	CN101482	Sel. of CIli-1676 (C4)	Oil	Breeding material	CAN
359	CN101486	Sel CIli -1761 (short)	Fiber	Breeding material	CAN
360	CN101493	Sel CIli -1819 (C4)	Oil	Breeding material	CAN
361	CN101496	Sel. of CIli-1856 (LS)	Oil	Breeding material	CAN
362	CN101510	Sel CIli -1966 (C4)	Oil	Breeding material	CAN
363	CN101511	Sel CIli -1967 (C5)	Oil	Breeding material	CAN
364	CN101535	Sel CIli -2085 (C2)	Oil	Breeding material	CAN
365	CN101536	Sel CIli -2085 (C4)	Oil	Breeding material	CAN
366	CN101539	Sel CIli -2155 (C4)	Oil	Breeding material	CAN
367	CN101542	Sel CIli -2197 (C4)	Oil	Breeding material	CAN
368	CN101554	Sel CIli -2225 (C4)	Oil	Breeding material	CAN

Sample	Accession number	Name	Type	Improvement status	Origin
369	CN101559	Sel CIli -2225 (C4)	Fiber	Breeding material	CAN
370	CN101560	Sel. of CIli-2289 (C2)	Oil	Breeding material	CAN
371	CN101565	Sel CIli -2410 (C6)	Oil	Breeding material	CAN
372	CN101572	Sel. of CIli-2560 (C4)	Unknown	Breeding material	CAN
373	CN101580	Sel. of CIli-2611 (C4)	Oil	Breeding material	CAN
374	CN101594	Sel. of CIli-2699 (C4)	Oil	Breeding material	CAN
375	CN101595	Sel CIli -2703 (C6/C4)	Oil	Breeding material	CAN
376	CN101596	Sel CIli -2719 (C4)	Oil	Breeding material	CAN
377	CN101598	Sel. of CIli-2734 (C4)	Oil	Breeding material	CAN
378	CN101600	Sel CIli -2748 (C4)	Oil	Breeding material	CAN
379	CN101610	Sel VIR-2404	Unknown	Breeding material	CAN
380	Linola989	Linola989	Oil	Cultivar	CAN
381	CDCGold	CDCGold	Oil	Cultivar	CAN
382	Macbeth	Macbeth	Oil	Cultivar	CAN
383	Shape	Shape	Oil	Cultivar	CAN
384	CDCSorrel	CDCSorrel	Oil	Cultivar	CAN
385	Atlas	Atlas	Fiber	Cultivar	SWE
386	CDCBethune	CDCBethune	Oil	Cultivar	CAN
387	CDCMons	CDCMons	Oil	Cultivar	CAN
388	CrepitamTabor	CrepitamTabor	Fiber	Cultivar	CAN
389	DoubleLow	DoubleLow	Oil	Cultivar	CAN
390	E1747	E1747	Oil	Cultivar	CAN
391	FP2214	FP2214	Oil	Cultivar	CAN

Sample	Accession number	Name	Type	Improvement status	Origin
392	FP2270	FP2270	Oil	Cultivar	CAN
393	G1186-94	G1186-94	Fiber	Cultivar	CAN
394	Hanley	Hanley	Oil	Cultivar	CAN
395	Lirina	Lirina	Oil	Cultivar	CAN
396	M5791	M5791	Oil	Cultivar	CAN
397	M96006	M96006	Oil	Cultivar	CAN
398	PrairieBlue	PrairieBlue	Oil	Cultivar	CAN
399	PrairieGrande	PrairieGrande	Oil	Cultivar	CAN
400	PrairieThunder	PrairieThunder	Oil	Cultivar	CAN
401	SP2047	SP2047	Oil	Cultivar	CAN
402	S95407	S95407	Oil	Cultivar	CAN
403	UGG102-2	UGG102-2	Oil	Cultivar	CAN
404	UGG146-1	UGG146-1	Oil	Cultivar	CAN
405	UGG5-5	UGG5-5	Oil	Cultivar	CAN
406	Viking(European)	Viking(European)	Fiber	Cultivar	EU
407	YSED18	YSED18	Oil	Cultivar	CAN



Appendix IV Distribution of the 407 flax accessions of the core collection. **a** geographical origin **b** improvement status.

Appendix V List of the 407 flax accessions sorted according to the neighbour-joining tree.

Number	Canadian number	Sub-group
120	**CN97605**	Western Europe
261	**CN100837**	Western Europe
274	**CN100910**	Western Europe
266	**CN100852**	Western Europe
272	**CN100885**	Western Europe
65	**CN97147**	Western Europe
368	**CN101554**	Western Europe
361	**CN101496**	Western Europe
269	**CN100881**	Western Europe
270	**CN100883**	Western Europe
199	**CN98535**	Western Europe
70	**CN97238**	Western Europe
219	**CN98741**	Western Europe
184	**CN98278**	Western Europe
46	**CN96911**	Western Europe
217	**CN98733**	Western Europe
218	**CN98734**	Western Europe
371	**CN101565**	Western Europe
216	**CN98712**	Western Europe
222	**CN98753**	Western Europe
375	**CN101595**	Western Europe
170	**CN98192**	Western Europe
171	**CN98193**	Western Europe
223	**CN98767**	Western Europe
114	**CN97571**	Western Europe
262	**CN100838**	Western Europe
77	**CN97321**	Western Europe
185	**CN98279**	Western Europe
304	**CN101265**	Western Europe
248	**CN100674**	Western Europe
71	**CN97287**	Western Europe
275	**CN100928**	Western Europe
280	**CN101026**	Western Europe
197	**CN98475**	Western Europe
235	**CN98923**	Western Europe
187	**CN98303**	Western Europe
388	**CrepitamTabor**	Western Europe
320	**CN101338**	South Asia
66	**CN97153**	South Asia
325	**CN101373**	South Asia

Number	Canadian number	Sub-group
168	**CN98165**	South Asia
48	**CN96962**	South Asia
256	**CN100799**	South Asia
62	**CN97129**	South Asia
260	**CN100828**	South Asia
63	**CN97129B**	South Asia
378	**CN101600**	South Asia
257	**CN100805**	South Asia
263	**CN100841**	South Asia
252	**CN100790**	South Asia
45	**CN96846**	South Asia
259	**CN100827**	South Asia
314	**CN101310**	South Asia
169	**CN98176**	South Asia
198	**CN98505**	South Asia
233	**CN98869**	South Asia
20	**CN19007**	South Asia
53	**CN97004**	South Asia
51	**CN96991**	South Asia
52	**CN96992**	South Asia
358	**CN101482**	South Asia
50	**CN96988**	South Asia
356	**CN101471**	South Asia
357	**CN101472**	South Asia
44	**CN96845**	South Asia
85	**CN97393**	South Asia
299	**CN101208**	South Asia
76	**CN97312**	South Asia
74	**CN97307**	South Asia
75	**CN97308**	South Asia
204	**CN98566C**	South Asia
202	**CN98566**	South Asia
203	**CN98566B**	South Asia
73	**CN97306**	South Asia
167	**CN98157**	South Asia
205	**CN98569**	South Asia
61	**CN97103**	South Asia
59	**CN97092**	South Asia
60	**CN97096**	South Asia
200	**CN98541**	South Asia
207	**CN98613**	South Asia
206	**CN98610**	South Asia

Number	Canadian number	Sub-group
376	**CN101596**	South Asia
244	**CN98982**	South Asia
364	**CN101535**	South Asia
365	**CN101536**	South Asia
366	**CN101539**	South Asia
195	**CN98467**	South Asia
196	**CN98468**	South Asia
194	**CN98440**	South Asia
47	**CN96958**	South Asia
174	**CN98239**	South Asia
247	**CN100629**	South Asia
58	**CN97083**	South Asia
54	**CN97050**	South Asia
362	**CN101510**	South Asia
363	**CN101511**	South Asia
178	**CN98250**	South Asia
56	**CN97064**	South Asia
55	**CN97056**	South Asia
188	**CN98363**	South Asia
180	**CN98263**	South Asia
181	**CN98263B**	South Asia
57	**CN97072**	South Asia
67	**CN97176**	South Asia
49	**CN96974**	South Asia
360	**CN101493**	South Asia
173	**CN98237**	South Asia
232	**CN98854**	South Asia
249	**CN100678**	South Asia
354	**CN101466**	South Asia
165	**CN98135**	South Asia
191	**CN98397**	South Asia
189	**CN98364**	South Asia
64	**CN97139**	South Asia
193	**CN98415**	South Asia
161	**CN98057**	South Asia
241	**CN98969**	South Asia
164	**CN98109**	South Asia
367	**CN101542**	South Asia
192	**CN98398**	South Asia
242	**CN98973**	South Asia
179	**CN98254**	South Asia
243	**CN98974**	South Asia

Number	Canadian number	Sub-group
190	**CN98370**	South Asia
177	**CN98242**	South Asia
240	**CN98961**	South Asia
175	**CN98240**	South Asia
176	**CN98240B**	South Asia
377	**CN101598**	South America
12	**CN18993**	South America
403	**UGG102-2**	South America
389	**DoubleLow**	South America
404	**UGG146-1**	South America
151	**CN97980**	South America
152	**CN98007**	South America
118	**CN97587**	South America
156	**CN98037**	South America
157	**CN98037B**	South America
150	**CN97967**	South America
158	**CN98039**	South America
149	**CN97961**	South America
147	**CN97953**	South America
148	**CN97958**	South America
212	**CN98689**	South America
124	**CN97633**	South America
208	**CN98634**	South America
163	**CN98100**	South America
95	**CN97430**	South America
96	**CN97430B**	South America
353	**CN101463**	South America
153	**CN98012**	South America
155	**CN98027**	South America
209	**CN98639**	North America-Europe
210	**CN98644**	North America-Europe
182	**CN98275**	North America-Europe
183	**CN98276**	North America-Europe
245	**CN98984**	North America-Europe
224	**CN98773**	North America-Europe
117	**CN97586**	North America-Europe
221	**CN98752**	North America-Europe
374	**CN101594**	North America-Europe
373	**CN101580**	North America-Europe
82	**CN97366**	North America-Europe
254	**CN100797**	North America-Europe
255	**CN100797B**	North America-Europe

Number	Canadian number	Sub-group
213	**CN98704**	North America-Europe
25	**CN19160**	North America-Europe
69	**CN97214**	North America-Europe
130	**CN97670**	North America-Europe
144	**CN97890**	North America-Europe
145	**CN97907**	North America-Europe
201	**CN98542**	North America-Europe
172	**CN98231**	North America-Europe
359	**CN101486**	North America-Europe
258	**CN100807**	North America-Europe
279	**CN101016**	North America-Europe
319	**CN101332**	North America-Europe
379	**CN101610**	North America-Europe
295	**CN101132**	North America-Europe
305	**CN101279**	North America-Europe
338	**CN101402**	North America-Europe
26	**CN30860**	North America-Europe
27	**CN30861**	North America-Europe
226	**CN98806**	North America-Europe
227	**CN98807**	North America-Europe
214	**CN98708**	North America-Europe
372	**CN101572**	North America-Europe
268	**CN100864**	North America-Europe
68	**CN97180**	North America-Europe
24	**CN19159**	North America-Europe
355	**CN101469**	North America-Europe
72	**CN97300**	North America-Europe
80	**CN97350**	North America-Europe
154	**CN98014**	North America-Europe
220	**CN98742**	North America-Europe
134	**CN97689**	North America
140	**CN97871**	North America
265	**CN100851**	North America
271	**CN100884**	North America
297	**CN101137**	North America
312	**CN101307**	North America
352	**CN101461**	North America
370	**CN101560**	North America
228	**CN98812**	North America
277	**CN100939**	North America
324	**CN101367**	North America
139	**CN97768**	North America

Number	Canadian number	Sub-group
133	**CN97679B**	North America
132	**CN97679**	North America
142	**CN97881**	North America
79	**CN97341**	North America
93	**CN97407**	North America
100	**CN97458**	North America
225	**CN98794**	North America
131	**CN97671**	North America
38	**CN33400**	North America
17	**CN19003**	North America
141	**CN97873**	North America
32	**CN33388**	North America
10	**CN18989**	North America
33	**CN33389**	North America
251	**CN100785**	North America
39	**CN33992**	North America
1	**CN18973**	North America
399	**PrairieGrande**	North America
400	**PrairieThunder**	North America
383	**Shape**	North America
382	**Macbeth**	North America
391	**FP2214**	North America
18	**CN19004**	North America
384	**CDCSorrel**	North America
344	**CN101413**	North America
4	**CN18981**	North America
19	**CN19005**	North America
30	**CN33385**	North America
3	**CN18980**	North America
43	**CN52732**	North America
21	**CN19017**	North America
28	**CN32542**	North America
250	**CN100770**	North America
78	**CN97334**	North America
135	**CN97718**	North America
137	**CN97740**	North America
329	**CN101382**	North America
136	**CN97728**	North America
138	**CN97749**	North America
146	**CN97921**	North America
229	**CN98821**	North America
127	**CN97642**	North America

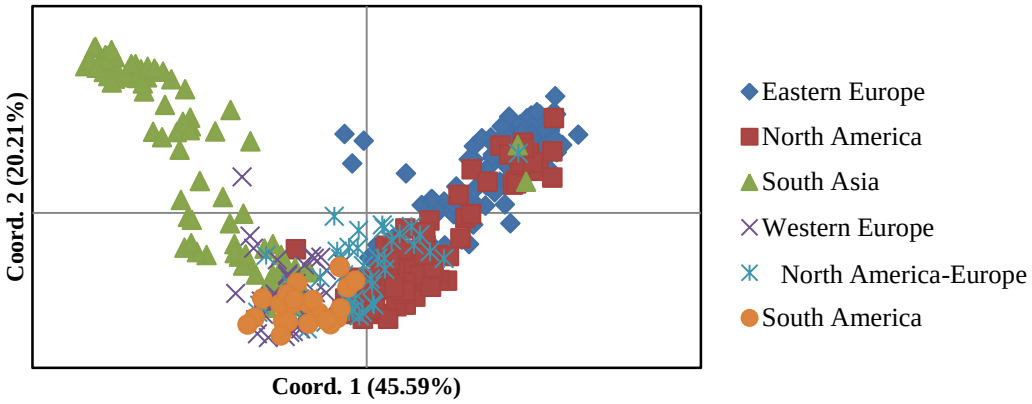
Number	Canadian number	Sub-group
37	**CN33399**	North America
395	**Lirina**	North America
331	**CN101386**	North America
13	**CN18994**	North America
36	**CN33397**	North America
392	**FP2270**	North America
386	**CDCBethune**	North America
387	**CDCMons**	North America
2	**CN18979**	North America
398	**PrairieBlue**	North America
396	**M5791**	North America
405	**UGG5-5**	North America
393	**G1186-94**	North America
402	**S95407**	North America
41	**CN37286**	North America
381	**CDCGold**	North America
407	**YSED18**	North America
380	**Linola989**	North America
401	**SP2047**	North America
394	**Hanley**	North America
390	**E1747**	North America
397	**M96006**	North America
323	**CN101366**	Eastern Europe
330	**CN101385**	Eastern Europe
98	**CN97452**	Eastern Europe
273	**CN100895B**	Eastern Europe
326	**CN101375**	Eastern Europe
84	**CN97392**	Eastern Europe
315	**CN101325**	Eastern Europe
318	**CN101331**	Eastern Europe
186	**CN98286**	Eastern Europe
122	**CN97613**	Eastern Europe
278	**CN100952**	Eastern Europe
246	**CN100547**	Eastern Europe
31	**CN33386**	Eastern Europe
385	**Atlas**	Eastern Europe
81	**CN97351**	Eastern Europe
336	**CN101397**	Eastern Europe
119	**CN97604**	Eastern Europe
92	**CN97406**	Eastern Europe
328	**CN101379**	Eastern Europe
316	**CN101327**	Eastern Europe

Number	Canadian number	Sub-group
321	**CN101348**	Eastern Europe
104	**CN97483**	Eastern Europe
128	**CN97649**	Eastern Europe
327	**CN101378**	Eastern Europe
83	**CN97377**	Eastern Europe
313	**CN101308**	Eastern Europe
87	**CN97397**	Eastern Europe
89	**CN97403**	Eastern Europe
99	**CN97453**	Eastern Europe
125	**CN97639**	Eastern Europe
126	**CN97639B**	Eastern Europe
86	**CN97396**	Eastern Europe
88	**CN97402**	Eastern Europe
90	**CN97404**	Eastern Europe
91	**CN97404B**	Eastern Europe
215	**CN98710**	Eastern Europe
237	**CN98934**	North America
306	**CN101286**	North America
350	**CN101451**	North America
22	**CN19157**	North America
267	**CN100863**	North America
349	**CN101448**	North America
94	**CN97424**	North America
369	**CN101559**	North America
236	**CN98926**	North America
239	**CN98954**	North America
211	**CN98683**	North America
238	**CN98946**	North America
162	**CN98072**	North America
23	**CN19158**	North America
264	**CN100848**	North America
351	**CN101454**	North America
115	**CN97584**	North America
116	**CN97584B**	North America
231	**CN98829**	North America
302	**CN101240**	Eastern Europe
105	**CN97484**	Eastern Europe
343	**CN101407**	Eastern Europe
110	**CN97529**	Eastern Europe
121	**CN97610**	Eastern Europe
111	**CN97530**	Eastern Europe
253	**CN100795**	Eastern Europe

Number	Canadian number	Sub-group
109	**CN97520**	Eastern Europe
123	**CN97616**	Eastern Europe
97	**CN97444**	Eastern Europe
101	**CN97463**	Eastern Europe
230	**CN98826**	Eastern Europe
106	**CN97487**	Eastern Europe
107	**CN97489**	Eastern Europe
113	**CN97533**	Eastern Europe
108	**CN97503**	Eastern Europe
129	**CN97665**	Eastern Europe
143	**CN97886**	Eastern Europe
322	**CN101364**	Eastern Europe
112	**CN97531**	Eastern Europe
102	**CN97470**	Eastern Europe
103	**CN97475**	Eastern Europe
234	**CN98903**	Eastern Europe
159	**CN98056**	Eastern Europe
160	**CN98056B**	Eastern Europe
166	**CN98150**	Eastern Europe
301	**CN101237**	Eastern Europe
332	**CN101392**	Eastern Europe
283	**CN101052**	Eastern Europe
284	**CN101053**	Eastern Europe
7	**CN18986**	Eastern Europe
406	**Viking(European)**	Eastern Europe
8	**CN18987**	Eastern Europe
5	**CN18982**	Eastern Europe
9	**CN18988**	Eastern Europe
341	**CN101405**	Eastern Europe
298	**CN101154**	Eastern Europe
293	**CN101119**	Eastern Europe
35	**CN33393**	Eastern Europe
276	**CN100929**	Eastern Europe
6	**CN18983**	Eastern Europe
11	**CN18991**	Eastern Europe
16	**CN19001**	Eastern Europe
15	**CN18998**	Eastern Europe
34	**CN33390**	Eastern Europe
42	**CN40081**	Eastern Europe
339	**CN101403**	Eastern Europe
317	**CN101329**	Eastern Europe
340	**CN101404**	Eastern Europe

Number	Canadian number	Sub-group
29	**CN32546**	Eastern Europe
296	**CN101136**	Eastern Europe
14	**CN18997**	Eastern Europe
307	**CN101289**	Eastern Europe
285	**CN101055**	Eastern Europe
40	**CN35791**	Eastern Europe
310	**CN101299**	Eastern Europe
342	**CN101406**	Eastern Europe
348	**CN101421**	Eastern Europe
345	**CN101416**	Eastern Europe
347	**CN101419**	Eastern Europe
300	**CN101230**	Eastern Europe
346	**CN101417**	Eastern Europe
281	**CN101038**	Eastern Europe
286	**CN101094**	Eastern Europe
291	**CN101116**	Eastern Europe
337	**CN101401**	Eastern Europe
292	**CN101118**	Eastern Europe
333	**CN101394**	Eastern Europe
287	**CN101096**	Eastern Europe
288	**CN101099**	Eastern Europe
309	**CN101298**	Eastern Europe
311	**CN101301**	Eastern Europe
282	**CN101039**	Eastern Europe
294	**CN101127**	Eastern Europe
335	**CN101396**	Eastern Europe
334	**CN101395**	Eastern Europe
308	**CN101296**	Eastern Europe
290	**CN101115**	Eastern Europe
289	**CN101114**	Eastern Europe
303	**CN101241**	Eastern Europe

a

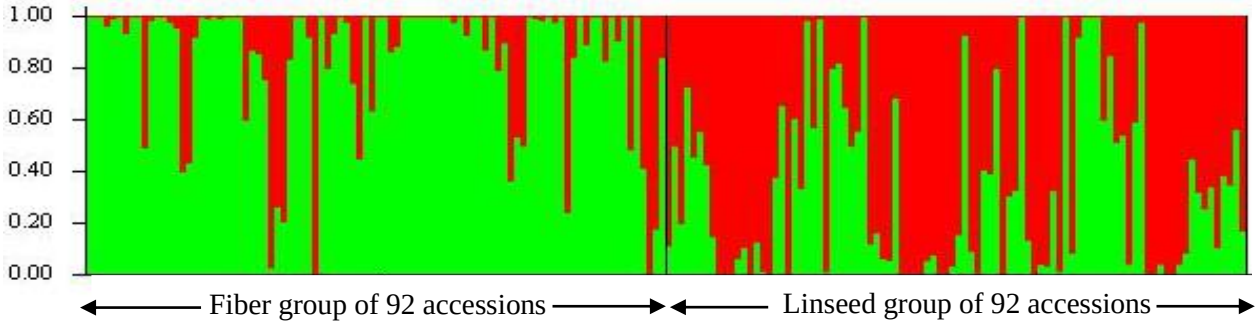


b

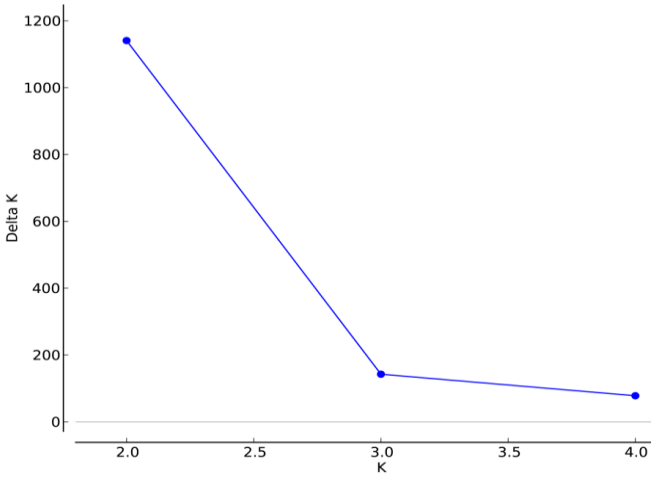
	1	2	3	4	5	6
1	—					
2	0.02*	—				
3	0.13*	0.16*	—			
4	0.08*	0.12*	0.07*	—		
5	0.04*	0.07*	0.09*	0.04*	—	
6	0.08*	0.12*	0.11*	0.05*	0.04*	—

Appendix VI Principal coordinate analysis (PCoA) and pairwise F_{ST} comparison between the sub-groups of flax **a** PCoA of the 407 flax accessions of the core collection based on the 259 neutral SSRs with LD < 0.4. Sub-groups were labeled according to the NJ analysis results (Fig. 3.1a) **b** Pairwise F_{ST} values between the 6 sub-groups of flax inferred by the NJ, STRUCTURE and PCoA analyses. 1 = North America. 2 = Eastern Europe. 3 = South Asia. 4 = Western Europe. 5 = North America/Europe. 6 = South America. * Significant values at P < 0.001.

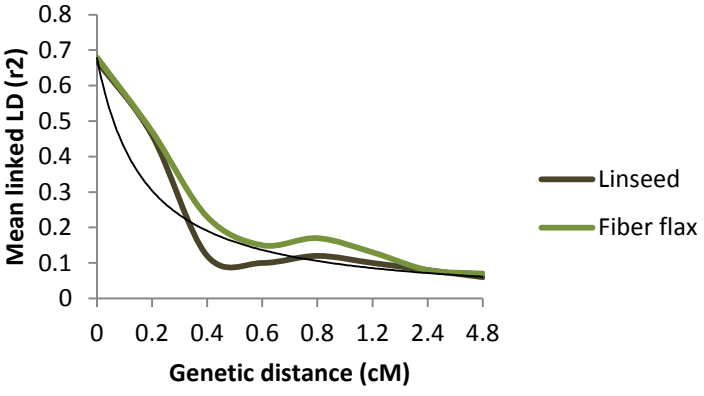
a



b



c



Appendix VII Population structure and linkage disequilibrium analyses of the fiber flax and linseed groups. **a** Bayesian clustering analysis (STRUCTURE $K = 2$) of fiber flax and linseed. **b** *ad-hoc* statistic ΔK Evanno et al. (2005) for K values ranging from 1 to 4. **c** Average genome-wide LD decay (r^2) against genetic distance (cM) within fiber and linseed flax groups. The black line represents the decay curve at the genome level of the two flax groups.

Appendix VIII Analysis of candidate genes affected by divergent selection between fiber flax and linseed groups.

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
1	s156-gene1-Lus10040693class=Sequence position=scaffold156:1042107..1044578 (+ strand	Full=50S ribosomal protein L4, chloroplastic;	165/219 (75%)	4e-106	O50061.2	GI:21542429	GO:0009570, GO:0009535, GO:0022626, GO:0005634, GO:0000311, GO:0008266, GO:0019843, GO:0003735,	*gb|JG259492 LUSTE1AD-RP-275_N17_20MAY2008_067 LUSTE1AD Linum usitatissimum cDNA, mRNA sequence Length = 869;Expect = 0.0 Identities = 823/826 (99%)*	
2	s156-gene-2-Lus10040694-class=Sequence position=scaffold156:1044953..1047480 (- strand)	Full=Phosphomethylpyrimidine synthase, chloroplastic;	261/278 (94%)	0	O82392.1	GI:75220243	GO:0051536, GO:0016829, GO:0046872, GO:0010266, GO:0009228, GO:0009229,	*gb|JG218635.1|JG218635 LUSST1AD-UP-101_E02_16JULY2008_008 LUSST1AD Linum usitatissimum cDNA, mRNA sequence Length = 779, Expect = 0.0, Identities = 719/721 (99%), Gaps = 1/721 (0%)*	
3	s156-Gene3-Lus10040695 class=Sequence position=scaffold156:1048891..1053637 (- strand)	Full=ATP-dependent zinc metalloprotease FTSH 7, chloroplastic;	139/239 (58%)	1e-63	Q6H6R9.1	GI:75323554	GO:0016021, GO:0005524, GO:0004222, GO:0017111, GO:0008270, GO:0030163, GO:0006508,	*genolin_c27713 338 nt similar to |AY149938|AY149938|Arabidopsis thaliana; At2g47010/F14M4.16 mRNA, complete cds; Length = 1960; Expect = e-118; Identities = 218/219 (99%)*	
4	s156-Gene4-Lus10040696 class=Sequence position=scaffold156:1064047..1065045 (+ strand)	Full=Transcription factor MYB1R1;	58/120 (48%)	8e-25	Q2V9B0.1	GI:122232932	GO:0005829, GO:0005634, GO:0003677, GO:0006355, GO:0006950, GO:0006351,	*genolin_c20781 195 nt Length = 1716; Expect = 0.0; Identities = 481/485 (99%)*	Cellulose microfibril angle, wood collapse in Eucaliptus piluraris (Sexton et al. 2011)
5	s156-Gene5-Lus10040697 class=Sequence position=scaffold156:1068694..1071500 (- strand)	Full=Benzenediol:oxygen oxidoreductase 22;	231/390 (59%)	0	Q0IQU1.2	GI:150383842	GO:0048046, GO:0005507, GO:0052716, GO:0046274, GO:0009834,	*gb|JG063140.1|JG063140 LUSES3AD-T3-001_D12_10AUG2009_046 LUSES1AD Linum usitatissimum cDNA, mRNA sequence Length = 855, Expect = 0.0, Identities = 487/488 (99%)*	
6	s291-Gene1-Lus10032405 class=Sequence	RecName: Full=Uncharacterized protein SLP1; AltName:	105/345 (30%)	2e-33	Q12232.1	GI:74676556	GO:0016021, GO:0034975,	*genolin_c34845 286 nt Length = 286 Expect = e-133 Identities = 277/286*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
	position=scaffold291 :1153041..1155600 (+ strand)	Full=SUN-like protein 1; Flags: Precursor Length=587						*(96%), Gaps = 2/286 (0%)*	
7	s291-Gene2-Lus10032406 class=Sequence position=scaffold291 :1156965..1157978 (+ strand)	Full=Cysteine endopeptidase; Flags: Precursor	195/342 (57%)	1e-125	O65039.1	GI:46395620	GO:0016023, GO:0008234, GO:0006508, GO:0005737, GO:0005524, GO:0004829, GO:0006435,	*gb|JG214310.1|JG214310 LUSPS1AD_RP_105_I10_ 14AUG2008_039 LUSPS1AD Linum usitatissimum cDNA, mRNA sequence Length = 856 Score = 58.0 bits (29), Expect = 9e-007 Identities = 62/73 (84%)*	expressed stem inner tissue (Fenart et al. 2010)
8	s291-Gene3-Lus10032407 class=Sequence position=scaffold291 :1159246..1161226 (+ strand)	Full=F-box protein SKIP8; AltName: Full=SKP1-interacting partner 8	53/71 (75%)	8e-30	Q93YV9.1	GI:75249436	GO:0016567,	*genolin_c35105 388 nt highly similar to |AL161518 Arabidopsis thaliana DNA chromosome 4, contig fragment No. 30 |Length = 738, Expect = e-163 Identities = 356/376 (94%), Gaps = 1/376 (0%)*	
9	s291-Gene4-Lus10032408 class=Sequence position=scaffold291 :1162354..1163193 (+ strand)	No BLASTx hit against UniProt	No Hits found					*gb|JG032954.1|JG032954 03-LUSBE1NG-RP-203_H12_31MAR2007_08 2 LUSBE1NG Linum usitatissimum cDNA, mRNA sequence Length = 653, Expect = e-157 Identities = 281/281 (100%)*	
10	s291-Gene5-Lus10032409 class=Sequence position=scaffold291 :1164172..1164444 (+ strand)	Full=RNA polymerase II transcriptional coactivator KELP Length=165	26/55 (47%)	3e-8	O65155.1	GI:37079408	GO:0005634, GO:0003677, GO:0003713, GO:0006355, GO:0006351,	*USHE1NG_RP_041_E05_ 09FEB2007_039.ab1; Length = 530; Expect = e-153; Identities = 273/273 (100%)*	
11	s291-Gene6-Lus10032410 class=Sequence position=scaffold291 :1165833..1167092 (+ strand)	Full=RNA polymerase II transcriptional coactivator KELP Length=165	75/171 (44%),	3e-26	O65155.1	GI:37079408	GO:0005634, GO:0003677, GO:0003713, GO:0006355, GO:0006351,	*gb|JG263915.1|JG263915 LUSTE1AD-RP-289_F05_22MAY2008_02 7 LUSTE1AD Linum usitatissimum cDNA, mRNA sequence Length = 817, Expect = e-168, Identities = 302/303 (99%)*	
12	s291-Gene7-	No BLASTx hit against						*genolin_c28591 232 nt*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
	Lus10032411 class=Sequence position=scaffold291 :1167483..1168082 (- strand)	UniProt						*Length = 797, Expect = 6e-096 Identities = 178/178 (100%)*	
13	s291-Gene8-Lus10032412 class=Sequence position=scaffold291 :1171667..1173288 (- strand)	Full=B2 protein Length=207	60/69 (87%	2e-34	P37707.1	GI:584825	No ontology	*gb|JG226758.1|JG226758 LUSST4AD-T3-041_M18_15SEP2009_067 LUSST1AD Linum usitatissimum cDNA, mRNA sequence Length = 792, Expect = 0.0, Identities = 430/432 (99%)*	
14	s917-Gene1-Lus10030641 class=Sequence position=scaffold917 :1149514..1149813 (- strand)	No BLASTx hit against UniProt						*genolin_c33219 350 nt, Length = 1666, Expect = 2e-048, Identities = 122/130 (93%)*	
15	s917-Gene2-Lus10030642 class=Sequence position=scaffold917 :1159087..1159546 (+ strand)	Full=Probable threonine--tRNA ligase, cytoplasmic; AltName: Full=Threonyl-tRNA synthetase; Short=ThrRS	28/52 (54%)	1e-7	Q8GZ45.2	GI:85701287	GO:0005737, GO:0005524, GO:0004829, GO:0006435,	*gb|CA482850.1|CA482850 LuP12001G08R LuP12 Linum usitatissimum cDNA clone LuP12001G08R, mRNA sequence, Length = 637, Expect = 2e-015, Identities = 106/127 (83%)*	
16	s917-Gene3-Lus10030643 class=Sequence position=scaffold917 :1160394..1162880 (- strand)	Full=Kinetochore protein NDC80 homolog; AltName: Full=Kinetochore protein Hec1; AltName: Full=Kinetochore-associated protein 2 Length=642	67/267 (25%)	8e-17	Q9D0F1.1	GI:81881154	GO:0000777, GO:0000942, GO:0031262, GO:0008608, GO:0051301, GO:0007059, GO:0000132, GO:0007067, GO:0007052,	*genolin_c14408 482 nt, Length = 1805, Expect = 0.0, Identities = 470/480 (97%), Gaps = 4/480 (0%)*	
17	s917-Gene4-Lus10030644 class=Sequence position=scaffold917 :1163157..1163716 (+ strand)	Full=Threonine--tRNA ligase, mitochondrial; AltName: Full=Threonyl-tRNA synthetase; Short=ThrRS; Flags: Precursor	42/75 (56%),	8e-26	O04630.3	GI:27735258	GO:0005618, GO:0009507, GO:0005829, GO:0005739, GO:0005886, GO:0005524, GO:0004829, GO:0006435,	*gb|JG151717.1|JG151717 LUSHE1NG-RP-076_F05_15FEB2007_037 LUSHE1NG Linum usitatissimum cDNA, mRNA sequence Length = 641, Expect = 3e-017, Identities = 103/122 (84%)*	
18	s917-Gene5-	RecName:	79/132	9e-86	O04630.3	GI:27735258	GO:0005618,	*genolin_c27668 527 nt*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
	Lus10030645 class=Sequence position=scaffold917 :1163873..1170020 (+ strand)	Full=Threonine--tRNA ligase, mitochondrial; AltName: Full=Threonyl-tRNA synthetase; Short=ThrRS; Flags: Precursor Length=709	(60%)				GO:0009507, GO:0005829, GO:0005739, GO:0005886, GO:0005524, GO:0004829, GO:0006435,	*Length = 1400, Expect = 1e-059 Identities = 167/182 (91%), Gaps = 2/182 (1%)*	
19	s208-Gene1-Lus10021711 class=Sequence position=scaffold208 :731243..733997 (+ strand)	RecName: Full=Transcription factor bHLH30; AltName: Full=Basic helix-loop-helix protein 30; Short=AtbHLH30; Short=bHLH 30; AltName: Full=Transcription factor EN 53; AltName: Full=bHLH transcription factor bHLH030 Length=368	101/173 (58%)	1e-34	Q9S7Y1.1	GI:75336852	GO:0005634, GO:0003677, GO:0006355, GO:0006351,	*gb|EB713935.1|EB713935 LuP12022C09R LuP12 Linum usitatissimum cDNA clone LuP12022C09, mRNA sequence Length = 449 Score = 531 bits (268), Expect = e-149 Identities = 275/276 (99%), Gaps = 1/276 (0%)*	
20	s208-Gene2-Lus10021712 class=Sequence position=scaffold208 :741204..742074 (- strand)	No BLASTx hit against UniProt						*gb|JG233667.1|JG233667 LUSTC1NG-RP-035_H07_27FEB2007_049 LUSTC1NG Linum usitatissimum cDNA, mRNA sequence Length = 626, Expect = 3e-034, Identities = 123/139 (88%)*	
21	s305-Gene1-Lus10025162 class=Sequence position=scaffold305 :591956..592495 (+ strand)	Full=RING-H2 finger protein ATL8 Length=185	74/117 (63%	8e-448	Q8LC69.2	GI:68565205	GO:0016021, GO:0008270, GO:0016567,	*gb|JG217382.1|JG217382 LUSPS1AD_RP_115_F02_ 18AUG2008_011 LUSPS1AD Linum usitatissimum cDNA, mRNA sequence Length = 905, Expect = 2e-030 Identities = 212/260 (81%)*	
22	s305-Gene2-Lus10025163 class=Sequence position=scaffold305 :596828..598803 (- strand)	Full=Probable inorganic phosphate transporter 1-9; Short=AtPht1;9; AltName: Full=H(+)/Pi cotransporter Length=532	161/307 (52%)	2e-77	Q9S735.1	GI:75313014	GO:0016021, GO:0015293, GO:0006817,	*LUSST3AD-T3-020_P17_5AUG2009_065, Length = 775, Expect = e-146, Identities = 263/263 (100%)*	
23	s305-Gene3-Lus10025164	Full=Probable inorganic phosphate transporter 1-	78/128 (61%)	7e-40	Q9SYQ1.2	GI:85687566	GO:0016021, GO:0015293,	*LUSST3AD-T3-020_P17_5AUG2009_065,*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
	class=Sequence position=scaffold305 :598835..600040 (+ strand)	8; Short=AtPht1;8; AltName: Full=H(+)/Pi cotransporter Length=534					GO:0006817	*Length = 775, Expect = e-170, Identities = 306/307 (99%)*	
24	s305-Gene4-Lus10025165 class=Sequence position=scaffold305 :601003..603766 (- strand)	RecName: Full=SWI/SNF complex subunit SWI3B; Short=AtSWI3B; AltName: Full=Transcription regulatory protein SWI3B Length=469	137/271 (51%)	6e-74	Q84JG2.1	GI:75327834	GO:0005634, GO:0003677, GO:0016568, GO:0007275, GO:0006355, GO:0006351,	*gb|JG134955.1|JG134955 LUSHE1AD-RP-281_B14_3JUNE2008_063 LUSHE1AD Linum usitatissimum cDNA, mRNA sequence Length = 569 Score = 482 bits (243), Expect = e-134 Identities = 243/243 (100%)*	
25	s225-Gene1-Lus10022341 class=Sequence position=scaffold225 :786295..790333 (+ strand)	No BLASTx hit against UniProt						*genolin_c24356 661 nt, Length = 3119, Expect = 0.0, Identities = 598/610 (98%)*	
26	s225-Gene2-Lus10022342 class=Sequence position=scaffold225 :790980..792191 (- strand)	Full=Probable adenylate kinase 1, chloroplastic; Short=AK 1; AltName: Full=ATP-AMP transphosphorylase 1; Flags: Precursor Length=284	56/151 (37%)	8e-19	Q9ZUU1.1	GI:29428074	GO:0005634, GO:0004017, GO:0005524, GO:0008652, GO:0048364, GO:0048367,	*gb|JG062715.1|JG062715 LUSES1AD_RP_103_F03_ 15JULY2008_012 LUSES1AD Linum usitatissimum, cDNA, mRNA sequence, Length = 729, Expect = 1e-021, Identities = 72/78 (92%)*	
27	s225-Gene3-Lus10022343 class=Sequence position=scaffold225 :800816..801259 (+ strand)	Full=Putative calcium-binding protein CML23; AltName: Full=Calmodulin-like protein 23 Length=151	81/150 (54%)	4e-42	Q8RYJ9.1	GI:75330796	GO:0005509	*genolin_c34622 211 nt, Length = 210, Expect = 5e-022, Identities = 54/54 (100%)*	
28	s225-Gene4-Lus10022344 class=Sequence position=scaffold225 :802650..803564 (- strand)	Full=DnaJ homolog subfamily C member 21; AltName: Full=DnaJ homolog subfamily A member 5	68/219 (31%)	1e-18	Q6PGY5.1	GI:82187285	GO:0005622, GO:0003676, GO:0008270, GO:0006457,	*LUSGE1NG_RP_120_C12 _09MAR2006_092.ab1, Length = 665, Expect = 0.0, Identities = 570/576 (98%)*	
29	s225-Gene5-Lus10022345 class=Sequence position=scaffold225	Full=Probable leucine-rich repeat receptor-like protein kinase At5g49770; Flags:	99/298 (33%)	2e-35	Q9LT96.1	GI:75335456	GO:0016021, GO:0005524, GO:0004674, GO:0004872,	*LUSGC1NG_RP_085_C09 _24JAN2007_075.ab1, Length = 725, Expect = e-135, Identities = 332/360*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
	:804542..806578 (-strand)	Precursor Length=946						*(92%), Gaps = 7/360 (1%)*	
30	s86-gene1-Lus10040449 class=Sequence position=scaffold86:1969440..1978828 (-strand)	Full=Uncharacterized WD repeat-containing protein alr2800 Length=1258	50/207 (24%)	3e-11	Q8YTC2.1	GI:20140995	GO:0043531, GO:0006952,	*gb|JG214541.1|JG214541 LUSPS1AD_RP_106_D12_14AUG2008_046 LUSPS1AD Linum usitatissimum cDNA, mRNA sequence, Length = 913, Expect = e-134, Identities = 253/256 (98%)*	
31	s86-Gene2-Lus10040450 class=Sequence position=scaffold86:1979512..1979772 (+ strand)	No BLASTx hit against UniProt	No Hits					*gb|JG285876.1|JG285876 LUSTE1NG-RP-179_C04_15FEB2007_028 LUSTE1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 423, Expect = e-146, Identities = 261/261 (100%)*	
32	s86-Gene3-Lus10040451 class=Sequence position=scaffold86:1980748..1982274 (+ strand)	Full=H/ACA ribonucleoprotein complex subunit 4; AltName: Full=CBF5 homolog; AltName: Full=Dyskerin; AltName: Full=Nopp-140-associated protein of 57 kDa homolog; Short=AtNAP57; AltName: Full=Nucleolar protein NAP57 homolog Length=565	287/326 (88%),	0	Q9LD90.1	GI:67460428	GO:0005829, GO:0005730, GO:0009506, GO:0030529, GO:0009982, GO:0003723, GO:0001522, GO:0006364,	*gb|JG241371.1|JG241371 LUSTC1NG-RP-128_A01_07MAR2007_015 LUSTC1NG Linum usitatissimum cDNA, mRNA sequence, Length = 749, Expect = 0.0, Identities = 745/749 (99%)*	
33	s86-Gene4-Lus10040452 class=Sequence position=scaffold86:1983902..1984927 (-strand)	Putative F-box/LRR-repeat protein At5g02930 Length=469	45/159 (28%),	3e-9	Q9LYZ2.1	GI:75264447	No ontology	*gb|JG247257.1|JG247257 LUSTC1NG-RP-199_B06_17MAR2007_046 LUSTC1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 633, Expect = 9e-004, Identities = 42/48 (87%)*	
34	s86-Gene5-Lus10040453 class=Sequence	Full=DnaJ homolog subfamily C member 2; AltName: Full=M-phase	23/52 (44%),	1e-5	Q99543.4	GI:296439472	GO:0005829, GO:0031965, GO:0003682,	No hit against flax ESTs	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
	position=scaffold86: 1987574..1987870 (- strand)	phosphoprotein 11; AltName: Full=Zuotin-related factor 1 Length=621					GO:0003677, GO:0051083, GO:0016568, GO:0006260, GO:0000085, GO:0030308, GO:0045893, GO:0006351,		
35	s280-Gene1-Lus10041365 class=Sequence position=scaffold280 :2125311..2128114 (+ strand)	No BLASTx hit against UniProt						gb|JG179335.1|JG179335 LUSLE4AD-T3-037_K22_06OCT2009_085 LUSLE1AD Linum usitatissimum cDNA, mRNA sequence, Length = 706, Expect = 0.0, Identities = 363/364 (99%)	
36	s280-Gene2-Lus10041366 class=Sequence position=scaffold280 :2129130..2133045 (- strand)	RecName: Full=Probable methyltransferase PMT15 Length=633 GENE ID: 825923 AT4G00750 | putative methyltransferase PMT15	96/133 (72%)	5e-54	Q9ZPH9.1	GI:75267756	GO:0005794, GO:0000139, GO:0016021, GO:0008168,	*genolin_c39764 236 nt Length = 794, Expect = e-129 Identities = 236/236 (100%)*	
37	s280-Gene3-Lus10041367 class=Sequence position=scaffold280 :2135272..2137021 (+ strand)	RecName: Full=F-box protein At4g00755 Length=377 GENE ID: 828014 AT4G00755 | F-box protein [Arabidopsis thaliana]	97/227	1e-42	Q8LG03.1	GI:75246091	No ontology	*gb|JG267291.1|JG267291 LUSTE1AD-RP-299_E04_27MAY2008_012 LUSTE1AD Linum usitatissimum cDNA, mRNA sequence, Expect = e-164, Identities = 293/293 (100%)*	
38	s280-Gene4-Lus10041368 class=Sequence position=scaffold280 :2137745..2138988 (- strand)	RecName: Full=Peroxisomal membrane protein 11C; AltName: Full=Peroxin-11C; Short=AtPEX11c Length=235	153/361 (42%)	6e-65	Q9LQ73.1	GI:75180079	GO:0005779, GO:0009506, GO:0016559,	*LUSBE1NG_RP_056_C05_25OCT2006_043.ab1, Length = 674, Expect = e-128, Identities = 306/328 (93%), Gaps = 2/328 (0%)*	
39	s280-Gene5-Lus10041369 class=Sequence position=scaffold280 :2140703..2142485 (- strand)	RecName: Full=Vacuolar protein 8 Length=556	33/133 (25%)	6e-7	Q5EFZ4.3	GI:74627608	GO:0005774, GO:0005488,	*gb|JG184357.1|JG184357 LUSME1AD-T3-011_J16_27JULY2009_056 LUSME1AD Linum usitatissimum, cDNA, mRNA sequence, Length =*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
								908, Expect = 0.0, Identities = 619/620 (99%)	
40	s280-Gene6-Lus10041370 class=Sequence position=scaffold280 :2143673..2144155 (+ strand)	Full=Galactinol--sucrose galactosyltransferase; AltName: Full=Raffinose synthase Length=798	38/83 (46%),	5e-16	Q8VWN6. 1	GI:75161213	GO:0047274, GO:0005975,	*genolin_c21095 396 nt, Length = 2251, Score = 228 bits (115), Expect = 2e-058, Identities = 220/253 (86%), Gaps = 9/253 (3%)*	
41	s280-Gene7-Lus10041371 class=Sequence position=scaffold280 :2147074..2147469 (- strand)	No BLASTx hit against UniProt						No hit against flax ESTs	
42	scaffold98_Gene1_3 97284_397652	No BLASTx hit against UniProt						*gb|JG129297.1|JG129297 LUSGE1NG-RP-351_E04_11JAN2008_024 LUSGE1NG Linum usitatissimum, cDNA, mRNA sequence, Length =588, Expect = 0.0, Identities = 369/369 (100%)*	
43	scaffold98_Gene2_3 99369_399713	No BLASTx hit against UniProt						*gb|JG041958.1|JG041958 03-LUSEN1NG-RP-042_H07_09MAR2007_04 9 LUSEN1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 459, Score = 676 bits (341), Expect = 0.0, Identities = 344/345 (99%)*	expressed stem outer tissue (Fenart et al. 2010)
44	scaffold98_Gene3_4 02350_402766	Full=Glutaredoxin-C9 Length=192	69/110 (63%),	5e-29	Q7XIZ1.1	GI:75142699	GO:0005737, GO:0005634, GO:0009055, GO:0015035, GO:0045454, GO:0022900, GO:0006810,	No hit against flax ESTs	
45	scaffold98_Gene4_4 06410_409511	Full=50S ribosomal protein L14	76/120 (63%)	1e-30	Q9ZCR5.1	GI:6225960	GO:0015934, GO:0019843, GO:0003735, GO:0006412,	*gb|JG019491.1|JG019491 03-LUSBE1NG-RP-022_E03_20OCT2006_023 LUSBE1NG Linum usitatissimum, cDNA,*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
								mRNA sequence, Length = 620, Expect = 0.0, Identities = 384/384 (100%)	
46	scaffold98_Gene5_4 14013_415454	RecName: Full=Transcription factor TCP2 Length=365	79/105 (75%),	4e-39	Q93V43.1	GI:75163104	GO:0005634, GO:0003677, GO:0030154, GO:0009965, GO:0045962, GO:0006355, GO:0006351,	*gb|JG226338.1|JG226338 LUSST4AD-T3-039_J14_15SEP2009_055 LUSST1AD Linum usitatissimum cDNA, mRNA sequence, Length = 900, Score = 1021 bits (515), Expect = 0.0, Identities = 524/527 (99%)*	
47	scaffold98_Gene6_4 22900_424174	Full=F-box protein At5g46170 Length=395 GENE ID: 834659 AT5G46170 | F-box protein [Arabidopsis thaliana]	272/384 (71%)	4e-161	Q93V43.1	GI:75163104	GO:0005634, GO:0003677, GO:0030154, GO:0009965, GO:0045962, GO:0006355, GO:0006351,	*genolin_c34571 411 nt, Length = 2067, Score = 541 bits (273), Expect = e-152, Identities = 276/279 (98%)*	
48	scaffold98_Gene7_4 25840_428477	Full=Probable ornithine aminotransferase; AltName: Full=Ornithine--oxo-acid aminotransferase Length=416	122/281 (43%)	2e-63	Q54JP5.1	GI:74896944	GO:0005737, GO:0004587, GO:0030170, GO:0006527, GO:0055129, GO:0006591,	*gb|JG108334.1|JG108334 LUSGE1NG-RP-080_A05_28FEB2007_047 LUSGE1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 663, Expect = e-122, Identities = 233/236 (98%)*	
49	scaffold98_Gene8_4 31313_435029	Full=KH domain-containing protein At4g18375 Length=606	109/200 (55%),	2e-52	P58223.1	GI:15214341	GO:0005634, GO:0003723,	*Genolin_c27804 465 nt similar to |AM435047|AM435047|Vit is vinifera contig, VV78X180421.8, whole genome shotgun sequence. Length = 465, Expect = 0.0, Identities = 389/405 (96%)*	
50	scaffold98_Gene9_4 36341_437072	No BLASTx hit against UniProt						*gb|JG077706.1|JG077706 LUSFL2AD-WB-010_H20_22NOV2008_07 4 LUSFL1AD Linum usitatissimum, cDNA, mRNA sequence, Length = 390, Expect = e-161,*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
								Identities = 287/287 (100%)	
51	scaffold98_Gene10_ 438516_440980	Full=Peroxisomal (S)-2-hydroxy-acid oxidase; AltName: Full=Glycolate oxidase; Short=GOX; AltName: Full=Short chain alpha-hydroxy acid oxidase Length=369	83/165 (50%)	7e-71	P05414.1	GI:121530	GO:0005777, GO:0010181, GO:0052853, GO:0052854, GO:0052852, GO:0009854,	*gb|JG091780.1|JG091780 LUSGC1NG-RP-131_E08_30JAN2007_056 LUSGC1NG Linum usitatissimum cDNA, mRNA sequence, Length = 489, Expect = 2e-099, Identities = 191/193 (98%)*	
52	scaffold98_Gene11_ 442330_442782	No BLASTx hit against UniProt						*gb|JG205486.1|JG205486 LUSME1NG-RP-189_A02_14APR2007_016 LUSME1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 534, Expect = 0.0, Identities = 412/421 (97%)*	
53	scaffold98_Gene12_ 446281_448331	RecName: Full=Pyruvate dehydrogenase E1 component subunit beta Length=326	250/322 (78%	4e-166	Q8MA03.1	GI:75272592	GO:0009507, GO:0004739, GO:0006096	*gb|JG256610.1|JG256610 LUSTE1AD-RP-266_D20_15MAY2008_078 LUSTE1AD Linum usitatissimum cDNA, mRNA sequence, Length = 792, Score = 1225 bits (618), Expect = 0.0, Identities = 621/622 (99%)*	
54	scaffold98_Gene13_ 449529_450067	No BLASTx hit against UniProt						*gb|CA483034.1|CA483034 LuP12003H04R LuP12 Linum usitatissimum cDNA clone LuP12003H04R, mRNA sequence, Length = 742, Expect = e-166, Identities = 323/332 (97%)*	
55	scaffold98_Gene14_ 450951_455179	RecName: Full=Cullin-4A; Short=CUL-4A Length=759	74/186 (40%),	1e-65	Q3TCH7.1	GI:108936014	GO:0031464, GO:0006281, GO:0044419, GO:0045732, GO:0045750, GO:0016567, GO:0006511	*gb|CA483034.1|CA483034 LuP12003H04R LuP12 Linum usitatissimum cDNA clone LuP12003H04R, mRNA sequence, Length = 742, Expect = e-137, Identities = 263/268 (98%)*	
56	scaffold98_Gene15_ 456575_456808	No BLASTx hit against UniProt						*gb|JG192842.1|JG192842 LUSME1NG-RP-040_F01_21FEB2007_005*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
								LUSME1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 477, Expect = e-130, Identities = 234/234 (100%)	
57	scaffold98_Gene16_460382_460720	No BLASTx hit against UniProt						*gb|JG088631.1|JG088631 LUSGC1NG-RP-094_D10_25JAN2007_074 LUSGC1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 463, Expect = 0.004, Identities = 22/22 (100%)*	
58	scaffold98_Gene17_465656_467484	Full=Glucan endo-1,3-beta-glucosidase 11; AltName: Full=(1->3)-beta-glucan endohydrolase 11; Short=(1->3)-beta-glucanase 11; AltName: Full=Beta-1,3-endoglucanase 11; Short=Beta-1,3-glucanase 11; Flags: Precursor Length=426	115/230 (50%),	3e-69	Q8L868.1	GI:75154301	GO:0046658, GO:0005618, GO:0005576, GO:0043169, GO:0042973, GO:0005975, GO:0007047, GO:0006952	*genolin_c42897 247 nt, Length = 862, Expect = e-128, Identities = 243/245 (99%), Gaps = 1/245 (0%)*	
59	scaffold98_Gene18_470801_476488	RecName: Full=Katanin p60 ATPase-containing subunit A-like 2; Short=Katanin p60 subunit A-like 2; AltName: Full=p60 katanin-like 2 Length=538	67/132 (51%)	5e-27	Q8IYT4.3	GI:189028467	GO:0005737, GO:0005874, GO:0005524, GO:0008568,	*gb|JG130726.1|JG130726 LUSGE1NG-RP-368_A06_17JAN2007_048 LUSGE1NG Linum usitatissimum cDNA, mRNA sequence, Length = 754, Expect = 0.0, Identities = 396/396 (100%)*	
60	scaffold98_Gene19_488842_490064	Full=Gibberellin 2-beta-dioxygenase; AltName: Full=GA 2-oxidase; AltName: Full=Gibberellin 2-beta-hydroxylase; AltName: Full=Gibberellin 2-oxidase Length=332	72/116 (62%)	7e-108	Q9XG83.1	GI:49035968	GO:0045543, GO:0005506, GO:0016702, GO:0009686,	*genolin_c40872 256 nt, Length = 2190, Expect = 8e-097, Identities = 180/180 (100%)*	
61	scaffold98_Gene20_	No BLASTx hit against						*gb|JG053799.1|JG053799*	expressed stem inner tissue

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
	499753_503369	UniProt						*03-LUSEN1NG-RP-179_H03_28MAR2007_017 LUSEN1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 648, Expect = e-180, Identities = 411/441 (93%)*	(Fenart et al. 2010)
62	scaffold98_Gene21_506403_508325	RecName: Full=Probable prefoldin subunit 6 Length=125	28/44 (64%)	1e-7	Q9VW56.1	GI:12230499	GO:0016272, GO:0006457,	*gb|JG281587.1|JG281587 LUSTE1NG-RP-130_H01_6FEB2006_001 LUSTE1NG Linum usitatissimum cDNA, mRNA sequence, Length = 671, Expect = 0.0, Identities = 667/671 (99%)*	
63	scaffold98_Gene22_509753_513251	No BLASTx hit against UniProt						*LUSHE1AD-RP-277_O04_30MAY2008_002, Length = 781, Expect = 0.0, Identities = 611/611 (100%)*	
64	scaffold98_Gene23_513885_516730	No BLASTx hit against UniProt						*gb|JG207296.1|JG207296 LUSME1NG-RP-209_C05_28NOV2007_043 LUSME1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 587, Expect = 1e-085, Identities = 201/214 (93%)*	
65	scaffold98_Gene24_518171_520057	Full=Phospho-N-acetylmuramoyl-pentapeptide-transferase homolog; AltName: Full=Translocase I Length=480	102/182 (56%)	1e-36	O49730.3	GI:229621258	GO:0016021, GO:0008963,	*gb|JG244959.1|JG244959 LUSTC1NG-RP-171_G04_17MAR2007_020 LUSTC1NG Linum usitatissimum cDNA, mRNA sequence, Length = 685, Expect = 8e-092, Identities = 202/212 (95%)*	
66	scaffold98_Gene25_525066_525488	Full=Probable purine permease 10; Short=AtPUP10 Length=390	73/126 (58%),	1e-42	O49725.2	GI:167012003	GO:0016021, GO:0016020, GO:0005345, GO:0009624,	No hit against flax ESTs	
67	scaffold98_Gene26_525551_526798	scaffold98_Gene26_525551_526798	51/81 (63%),	7e-21	O49725.2	GI:167012003	GO:0016021, GO:0016020, GO:0005345, GO:0009624,	*genolin_c13306 513 nt, Length = 511, Expect = 0.0, Identities = 328/328 (100%)*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
68	scaffold98_Gene27_528586_533646	Full=Glycogen synthase 2; AltName: Full=Starch [bacterial glycogen] synthase 2 Length=487	66/159 (42%)	2e-30	Q604D9.2	GI:91206712	GO:0009011, GO:0005978,	*genolin_c30229 306 nt, Length = 1022, Expect = 0.0, Identities = 671/676 (99%)*	
69	scaffold98_Gene28_533921_534217	Full=Chlorophyll a-b binding protein 1D; AltName: Full=LHCII type I CAB-1D; Short=LHCP Length=116	92/98 (94%),	1e-64	P10707.1	GI:115822	GO:0009535, GO:0016021, GO:0009522, GO:0009523, GO:0016168, GO:0046872, GO:0009765, GO:0018298,	*gb|JG222986.1|JG222986 LUSST4AD-T3-020_O01_15SEP2009_001 LUSST1AD Linum usitatissimum, cDNA, mRNA sequence, Length = 744*	
70	scaffold98_Gene29_534348_534719	Chlorophyll a-b binding protein 16, chloroplastic; AltName: Full=LHCII type I CAB-16; Short=LHCP; Flags: Precursor Length=266	82/103 (80%	2e-54	P27492.1	GI:115781	GO:0009535, GO:0016021, GO:0009522, GO:0009523, GO:0016168, GO:0046872, GO:0009765, GO:0018298,	*gb|JG212961.1|JG212961 LUSPS1AD_RP_101_H13_11AUG2008_057 LUSPS1AD Linum usitatissimum, cDNA, mRNA sequence, Length = 868, Expect = 0.0, Identities = 362/372 (97%), Gaps = 1/372 (0%)*	
71	scaffold98_Gene30_536533_537144	Full=Chlorophyll a-b binding protein 3C, chloroplastic; AltName: Full=LHCII type I CAB-3C; Short=LHCP; Flags: Precursor Length=267	178/266 (67%)	1e-114	P07369.1	GI:115825	GO:0009535, GO:0016021, GO:0009522, GO:0009523, GO:0016168, GO:0046872, GO:0009765, GO:0018298,	*gb|JG222967.1|JG222967 LUSST4AD-T3-020_M07_15SEP2009_020 LUSST1AD Linum usitatissimum, cDNA, mRNA sequence, Length = 819, Expect = 0.0, Identities = 365/365 (100%)*	
72	scaffold98_Gene31_537930_542065	Full=LRR receptor-like serine/threonine-protein kinase FLS2; AltName: Full=Protein FLAGELLIN-SENSING 2; AltName: Full=Protein FLAGELLIN-SENSITIVE 2; Flags: Precursor Length=1173	473/945 (50%)	0	Q9FL28.1	GI:75262640	GO:0010008, GO:0016021, GO:0005886, GO:0005524, GO:0004674, GO:0004872, GO:0052544, GO:0042742, GO:0016045, GO:0010359,	*genolin_c28812 266 nt, Length = 1422, Expect = e-140, Identities = 263/265 (99%), Gaps = 1/265 (0%)*	
73	scaffold98_Gene32_544581_544823	RecName: Full=Probable receptor-like protein kinase At5g38990; Flags:	25/70 (36%)	3e-06	Q9FID9.1	GI:75333907	GO:0016021, GO:0005524, GO:0004674,	*genolin_c12301 193 nt, Length = 3746, Expect = 1e-008, Identities = 58/67 (86%)*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
		Precursor Length=880							
74	scaffold98_Gene33_545289_545978	Receptor-like protein kinase HERK 1; AltName: Full=Protein HERCULES RECEPTOR KINASE 1; Flags: Precursor Length=830	82/225 (36%)	2e-25	Q9LX66.1	GI:75335601	GO:0016021, GO:0005886, GO:0009506, GO:0005524, GO:0004672, GO:0004674, GO:0004872, GO:0009742, GO:0009791, GO:0051510, GO:0009826,	*genolin_c12301 193 nt, Length = 3746, Expect = 4e-008, Identities = 34/35 (97%)*	
75	scaffold98_Gene34_546962_548755	Full=Polyneuridine-aldehyde esterase; AltName: Full=Polyneuridine aldehyde esterase; Flags: Precursor Length=264	57/108 (53%),	1e-28	Q9SE93.1	GI:50401192	GO:0004091, GO:0050529, GO:0009820	*gb|JG225691.1|JG225691 LUSST4AD-T3-036_C21_15SEP2009_093 LUSST1AD Linum usitatissimum cDNA, mRNA sequence, Length = 809, Expect = e-167, Identities = 298/298 (100%)*	
76	scaffold98_Gene35_549762_550229	No BLASTx hit against UniProt						*gb|JG101512.1|JG101512 LUSGC1NG-RP-245_F08_02APR2007_054 LUSGC1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 563, Score = 77.8 bits (39), Expect = 4e-013, Identities = 54/59 (91%)*	
77	scaffold98_Gene36_553003_560612	Full=Calcium-transporting ATPase 1, endoplasmic reticulum-type Length=1061	388/467 (83%)	0	P92939.2	GI:12643704	GO:0030176, GO:0005886, GO:0005774, GO:0005524, GO:0005388, GO:0046872, GO:0030026, GO:0006828, GO:0046686, GO:0010042	*genolin_c16266 441 nt, Length = 1851, Expect = 0.0, Identities = 736/758 (97%), Gaps = 3/758 (0%)*	
78	scaffold98_Gene37_561231_562778	Full=TMV resistance protein N Length=1144	56/182 (31%),	5e-28	Q40392.1	GI:46577339	GO:0005737, GO:0043531, GO:0005524, GO:0009626,	*gb|JG250981.1|JG250981 LUSTC1NG-RP-243_D11_15MAY2007_089 LUSTC1NG Linum*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
							GO:0007165	*usitatissimum cDNA, mRNA sequence Length = 661, Expect = 0.001, Identities = 33/36 (91%)*	
79	scaffold98_Gene38_563086_563918	RecName: Full=Leucine-rich repeat-containing protein 40 Length=602	36/120 (30%)	4e-8	Q9H9A6.1	GI:74761553	No ontology	No hit against flax ESTs	
80	scaffold98_Gene39_566883_567194	Full=Histone H4 variant TH091 Length=103	82/82 (100%)	5e-51	P62786.2	GI:51338727	GO:0000786, GO:0005634, GO:0003677, GO:0006334,	*gb|JG130751.1|JG130751 LUSGE1NG-RP-368_C11_17JAN2007_091 LUSGE1NG Linum usitatissimum, cDNA, mRNA sequence Length = 614, Score = 618 bits (312), Expect = e-176, Identities = 312/312 (100%)*	
81	scaffold98_Gene40_572911_573630	Full=WUSCHEL-related homeobox 3; AltName: Full=Protein PRESSED FLOWER Length=244	47/57 (82%)	3e-25	Q9SIB4.1	GI:61217434	GO:0005634, GO:0043565, GO:0003700, GO:0009943, GO:0030154, GO:0008283, GO:0009947, GO:0009908, GO:0010865, GO:0006351,	*gb|JG214412.1|JG214412 LUSPS1AD_RP_105_N07_14AUG2008_020 LUSPS1AD Linum usitatissimum, cDNA, mRNA sequence Length = 898, Expect = 6e-010, Identities = 49/54 (90%)*	
82	scaffold98_Gene41_579841_580740	Full=Chitinase 1; AltName: Full=Tulip bulb chitinase-1; Short=TBC-1; Flags: Precursor Length=314	165/278 (59%)	1e-112	Q9SLP4.1	GI:47605559	GO:0043169, GO:0008061, GO:0004568, GO:0006032,	*gb|JG106882.1|JG106882 LUSGE1NG-RP-063_C11_23FEB2007_091 LUSGE1NG Linum usitatissimum, cDNA, mRNA sequence Length = 692, Expect = 0.0, Identities = 325/326 (99%)*	
83	scaffold98_Gene42_582604_583287	No BLASTx hit against UniProt					No Hit	*LUSGE1NG_RP_191_C10_04APR2007_076.ab1, Length = 727, Expect = 0.0, Identities = 561/561 (100%)*	
84	scaffold98_Gene43_583631_584381	Full=Receptor-like serine/threonine-protein	58/140 (41%),	8e-34	O81905.1	GI:75318808	GO:0016021, GO:0005886,	*genolin_c35587 636 nt, Length = 635, Score = 165*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
		kinase SD1-8; AltName: Full=Arabidopsis thaliana receptor kinase 3; AltName: Full=S-domain-1 (SD1) receptor kinase 8; Short=SD1-8; Flags: Precursor Length=850					GO:0009506, GO:0005773, GO:0005524, GO:0004674, GO:0004872, GO:0048544,	*bits (83), Expect = 4e-039, Identities = 119/131 (90%)*	
85	scaffold98_Gene44_585760_588145	RecName: Full=BTB/POZ domain-containing protein At5g60050 Length=499	193/322 (60%),	3e-110	Q9LVG9.1	GI:75180651	No ontology	*gb|EH792489.1|EH792489 LU01UID.9374 stem phloem (bast) fiber enriched library LU01 Linum usitatissimum cDNA clone FLAXPH19_UP_001_D05, mRNA sequence Length = 1246 Expect = 0.0 dentities = 799/820 (97%), Gaps = 12/820 (1%)*	expressed stem outer tissue (Fenart et al. 2010)
86	scaffold98_Gene45_589675_592030	Full=ATP synthase subunit d, mitochondrial; Short=ATPase subunit d Length=168	45/57 (79%),	9e-22	Q9FT52.3	GI:25089786	GO:0009535, GO:0022626, GO:0005753, GO:0000276, GO:0005730, GO:0005774, GO:0005507, GO:0015078, GO:0016787, GO:0008270, GO:0015986, GO:0009651,	*gb|JG266257.1|JG266257 LUSTE1AD-RP-296_D04_26MAY2008_01 4 LUSTE1AD Linum usitatissimum cDNA, mRNA sequence Length = 846, Expect = 2e-092, Identities = 173/173 (100%)*	
87	scaffold98_Gene46_593212_595545	No BLASTx hit against UniProt						*gb|JG227288.1|JG227288 LUSST4AD-T3-044_N02_15SEP2009_003 LUSST1AD Linum usitatissimum cDNA, mRNA sequence Length = 914, Expect = e-172, Identities = 343/355 (96%)*	
88	scaffold98_Gene47_598919_602248	Full=Mitogen-activated protein kinase 3; Short=AtMPK3; Short=MAP kinase 3 Length=370	97/114 (85%)	2e-57	Q39023.2	GI:21431794	GO:0005737, GO:0005634, GO:0005524, GO:0004707, GO:0004672,	*gb|JG143087.1|JG143087 LUSHE1AD-RP-307_M16_16JUN2008_05 2 LUSHE1AD Linum usitatissimum cDNA,*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
							GO:0009738, GO:0000169, GO:0010120, GO:0000165, GO:0048481, GO:0009626, GO:0080136, GO:2000038, GO:2000037, GO:0009617, GO:0010200, GO:0009409, GO:0006970, GO:0006979, GO:0010224, GO:0009611,	*mRNA sequence, Length = 767, Expect = 2e-010, Identities = 87/104 (83%)*	
89	scaffold98_Gene48_604947_610070	Full=Non-lysosomal glucosylceramidase; Short=NLGase; AltName: Full=Beta-glucocerebrosidase 2; Short=Beta-glucosidase 2; AltName: Full=Glucosylceramidase 2 Length=927	66/117 (56%)	2e-29	Q9HCG7.2	GI:143018392	GO:0016021, GO:0005792, GO:0005886, GO:0005790, GO:0008422, GO:0004348, GO:0008206, GO:0006680, GO:0016139, GO:0006687,	*gb|JG240063.1|JG240063 LUSTC1NG-RP-112_D08_06MAR2007_058 LUSTC1NG Linum usitatissimum cDNA, mRNA sequence, Length = 704, Expect = e-153, Identities = 278/279 (99%)*	
90	scaffold98_Gene49_611633_615677	RecName: Full=DNA repair protein RAD51 homolog 2; Short=AtRAD51B Length=370	78/173 (45%)	1e-19	Q9SK02.2	GI:83305358	GO:0005634, GO:0005524, GO:0003677, GO:0008094, GO:0006310, GO:0006281,	No hit against flax ESTs	
91	scaffold98_Gene50_616904_617501	No BLASTx hit against UniProt						*gb|JG124790.1|JG124790 LUSGE1NG-RP-299_B02_23NOV2007_014 LUSGE1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 498, Expect = e-104, Identities = 207/212 (97%)*	
92	scaffold98_Gene51_617938_619130	No BLASTx hit against UniProt						*genolin_c17438 413 nt, Length = 1641, Expect = 4e-006, Identities = 31/32 (96%)*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
93	scaffold98_Gene52_621335_628038	Full=E3 SUMO-protein ligase SIZ1 Length=884	163/367 (44%)	2e-89	Q680Q4.2	GI:73919315	GO:0016607, GO:0005634, GO:0003676, GO:0019789, GO:0008270, GO:0051301, GO:0016049, GO:0016036, GO:0006952, GO:0010247, GO:0048589, GO:0009908, GO:0010286, GO:0009910, GO:0010113, GO:0016925, GO:0009787, GO:0040008, GO:0090352, GO:2000070, GO:0050826, GO:0009414, GO:0010337,	*genolin_c41078 239 nt, Length = 239, Expect = e-125, Identities = 235/237 (99%)*	
94	scaffold98_Gene53_628666_637849	No BLASTx hit against UniProt						*gb|JG219065.1|JG219065 LUSST3AD-T3-013_F09_5AUG2009_043 LUSST1AD Linum usitatissimum cDNA, mRNA sequence, Length = 744, Expect = 0.0, Identities = 737/741 (99%)*	
95	scaffold98_Gene54_639158_643411	Full=Calcium-dependent protein kinase 17 Length=528	160/172 (93%),	5e-99	Q9FMP5.1	GI:75334077	GO:0005737, GO:0005886, GO:0005524, GO:0005509, GO:0004674, GO:0046777, GO:0080092,	*gb|JG074943.1|JG074943 LUSFL1AD-WB-009_H11_09NOV2008_042 LUSFL1AD Linum usitatissimum cDNA, mRNA sequence Length = 934, Expect = 9e-081, Identities = 154/154 (100%)*	
96	scaffold98_Gene55_645707_646558	No BLASTx hit against UniProt						*gb|JG189528.1|JG189528 LUSME2AD-T3-023_B10_6AUG2009_047 LUSME1AD Linum*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
								usitatissimum cDNA, mRNA sequence, Length = 766, Expect = 2e-004, Identities = 28/29 (96%)	
97	scaffold98_Gene56_649787_652129	Full=Tubulin beta-5 chain; AltName: Full=Beta-5-tubulin Length=447	206/210 (98%)	2e-139	P46265.1	GI:1174600	GO:0005618, GO:0009507, GO:0005874, GO:0005886, GO:0005525, GO:0003924, GO:0005198, GO:0007018, GO:0051258, GO:0046686,	*gb|JG075806.1|JG075806 LUSFL1AD-WB-024_I20_10NOV2008_072 LUSFL1AD Linum usitatissimum cDNA, mRNA sequence, Length = 924, Expect = 0.0, Identities = 664/666 (99%)*	
98	scaffold98_Gene57_652941_653766	Full=Putative germin-like protein 2-1; Flags: Precursor Length=216	85/176 (48%)	7e-54	Q6K5Q0.1	GI:75261355	GO:0048046, GO:0005618, GO:0031012, GO:0030145, GO:0045735, GO:0009651,	*gb|JG232323.1|JG232323 LUSTC1NG-RP-020_A05_26FEB2007_047 LUSTC1NG Linum usitatissimum, cDNA, mRNA sequence Length = 579, Expect = 4e-073, Identities = 225/252 (89%), Gaps = 1/252 (0%)*	
99	scaffold98_Gene58_655609_656369	Full=Putative germin-like protein 2-1; Flags: Precursor Length=216	81/175 (46%	4e-52	Q6K5Q0.1	GI:75261355	GO:0048046, GO:0005618, GO:0031012, GO:0030145, GO:0045735, GO:0009651,	*gb|JG232323.1|JG232323 LUSTC1NG-RP-020_A05_26FEB2007_047 LUSTC1NG Linum usitatissimum, cDNA, mRNA sequence Length = 579, Expect = e-103, Identities = 240/255 (94%), Gaps = 1/255 (0%)*	
100	scaffold98_Gene59_660073_660723	Full=Putative germin-like protein 2-1; Flags: Precursor Length=216	131/216 (61%)	2e-88	Q6K5Q0.1	GI:75261355	GO:0048046, GO:0005618, GO:0031012, GO:0030145, GO:0045735, GO:0009651,	*gb|JG214720.1|JG214720 LUSPS1AD_RP_106_M16_14AUG2008_052 LUSPS1AD Linum usitatissimum, cDNA, mRNA sequence, Length = 879, Score = 56.0 bits (28), Expect = 2e-006, Identities = 34/36 (94%)*	
101	scaffold98_Gene60_661505_663492	Full=Probable protein phosphatase 2C 28; Short=AtPP2C28	23/35 (66%)	8e-6	O64583.2	GI:391358160	GO:0046872, GO:0004721, GO:0008152,	*gb|JG193905.1|JG193905 LUSME1NG-RP-052_B05_21FEB2007_045*	expressed stem inner tissue (Fenart et al. 2010)

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
		Length=339						*LUSME1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 531, Expect = 2e-061, Identities = 145/153 (94%)*	
102	scaffold98_Gene61_ 664643_671706	Full=PHD finger-containing protein DDB_G0268158 Length=688	50/165 (30%)	2e-6	Q55FD6.1	GI:74859221	GO:0008270,	*b|JG224385.1|JG224385 LUSST4AD-T3-028_I01_15SEP2009_007 LUSST1AD Linum usitatissimum, cDNA, mRNA sequence, Length = 888, Expect = 0.0, Identities = 708/709 (99%)*	expressed stem inner tissue (Fenart et al. 2010)
103	scaffold98_Gene62_ 673162_680563	No BLASTx hit against UniProt						*genolin_c11483 690 nt, Length = 2315, Expect = 0.0, Identities = 476/479 (99%)*	expressed stem inner tissue (Fenart et al. 2010)
104	scaffold98_Gene63_ 685661_687664	Full=DNA-damage-repair/toleration protein DRT111, chloroplastic; Flags: Precursor Length=387	186/292 (64%)	2e-58	P42698.2	GI:20141383	GO:0009507, GO:0005737, GO:0005634, GO:0000166, GO:0003723, GO:0006281,	*genolin_c36779 288 nt, Length = 288, Score = 93.7 bits (47), Expect = 3e-017, Identities = 116/139 (83%)*	
105	scaffold98_Gene64_ 688645_689527	No BLASTx hit against UniProt						*gb|JG132151.1|JG132151 LUSHE1AD-RP-272_I07_30MAY2008_024 LUSHE1AD Linum usitatissimum, cDNA, mRNA sequence, Length = 802, Expect = 8e-081, Identities = 153/153 (100%)*	
106	scaffold98_Gene65_ 692597_697218	Full=Homeobox-leucine zipper protein HOX32; AltName: Full=HD-ZIP protein HOX32; AltName: Full=Homeodomain transcription factor HOX32; AltName: Full=OsHox32 Length=859	134/245 (55%)	0	Q6AST1.1	GI:75119691	GO:0005634, GO:0043565, GO:0003700, GO:0006351,	*gb|JG179263.1|JG179263 LUSLE4AD-T3-037_F16_06OCT2009_060 LUSLE1AD Linum usitatissimum, cDNA, mRNA sequence, Length = 931, Score = 660 bits (333), Expect = 0.0, Identities = 333/333 (100%)*	
107	scaffold98_Gene66_ 699934_700982	Full=Nuclear transcription factor Y	48/58 (83%)	2e-23	Q84JP1.1	GI:75146690	GO:0005634, GO:0003677,	*gb|JG214921.1|JG214921 LUSPS1AD_RP_107_G14*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
		subunit A-7; Short=AtNF-YA-7 Length=190					GO:0003700, GO:0045892, GO:0006351,	*_15AUG2008_057 LUSPS1AD Linum usitatissimum, cDNA, mRNA sequence, Length = 890, Expect = e-106, Identities = 196/196 (100%)*	
108	scaffold98_Gene67_ 705511_706715	Full=Fatty acid 2-hydroxylase; AltName: Full=Fatty acid alpha-hydroxylase Length=372	47/143 (33%),	7e-36	Q2LAM0. 2	GI:162416308	GO:0005789, GO:0016021, GO:0005792, GO:0020037, GO:0016491, GO:0022900, GO:0006633, GO:0006665, GO:0006810,	*gb|JG265935.1|JG265935 LUSTE1AD-RP-295_E05_26MAY2008_02 7 LUSTE1AD Linum usitatissimum, cDNA, mRNA sequence, Length = 723, Expect = 3e-078, Identities = 149/149 (100%)*	
109	scaffold98_Gene68_ 708381_714139	Full=Myosin-2 heavy chain; AltName: Full=Myosin II heavy chain Length=2116	106/479 (22%)	7e-11	P08799.3	GI:134047850	GO:0042641, GO:0005826, GO:0032009, GO:0032982, GO:0016460, GO:0001931, GO:0030898, GO:0005524, GO:0000146, GO:0033275, GO:0032060, GO:0006935, GO:0030038, GO:0030866, GO:0031154, GO:0000910, GO:0060328, GO:0046847, GO:0031034, GO:0030837, GO:0008104, GO:0031270, GO:0034461	*LUSPS1AD_RP_104_H24 _14AUG2008_090, Length = 919, Expect = e-173, Identities = 310/310 (100%)*	expressed stem outer tissue (Fenart et al. 2010)
110	scaffold98_Gene69_ 723202_724905	Full=Reticuline oxidase-like protein; Flags: Precursor Length=570	227/538 (42%)	1e-135	Q9SVG4.2	GI:118585329	GO:0031225, GO:0048046, GO:0005829, GO:0005739, GO:0009505,	*LUSTC1NG_RP_153_B08 _12MAR2007_062.ab1, Length = 715, Expect = 0.0, Identities = 549/549 (100%)*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
							GO:0005886, GO:0009506, GO:0005773, GO:0050660, GO:0008762, GO:0006979,		
111	scaffold98_Gene70_735327_736578	No BLASTx hit against UniProt						*genolin_c19393 422 nt, Length = 1037, Expect = 0.004, Identities = 53/63 (84%)*	
112	scaffold98_Gene71_736742_738728	Full=Reticuline oxidase-like protein; Flags: Precursor Length=570	165/361 (46%)	6e-90	Q9SVG4.2	GI:118585329	GO:0031225, GO:0048046, GO:0005829, GO:0005739, GO:0009505, GO:0005886, GO:0009506, GO:0005773, GO:0050660, GO:0008762, GO:0006979	*genolin_c35055 364 nt, Length = 1078, Expect = 7e-009, Identities = 39/41 (95%)*	
113	scaffold98_Gene72_741160_741579	Full=Polygalacturonase; Short=PG; AltName: Full=Pectinase; Flags: Precursor Length=396	63/113 (56%),	8e-35	Q05967.1	GI:548491	GO:0005618, GO:0005576, GO:0004650, GO:0005975, GO:0007047	*gb|JG181357.1|JG181357 LUSLE4AD-T3-052_I23_13NOV2009_088 LUSLE1AD Linum usitatissimum, cDNA, mRNA sequence, Length = 934, Expect = 0.005, Identities = 37/42 (88%)*	
114	scaffold98_Gene73_741731_743788	Full=Reticuline oxidase-like protein; Flags: Precursor Length=570	200/443 (45%)	3e-103	Q9SVG4.2	GI:118585329	GO:0031225, GO:0048046, GO:0005829, GO:0005739, GO:0009505, GO:0005886, GO:0009506, GO:0005773, GO:0050660, GO:0008762, GO:0006979,	*genolin_c21177 543 nt, Length = 1581, Expect = 3e-005, Identities = 99/123 (80%)*	
115	scaffold98_Gene74_745613_746839	Full=Reticuline oxidase-like protein; Flags: Precursor Length=570	248/407 (61%)	1e-159	Q9SVG4.2	GI:118585329	GO:0031225, GO:0048046, GO:0005829, GO:0005739,	*genolin_c16532 283 nt, Length = 895, Expect = 0.0, Identities = 539/558 (96%), Gaps = 3/558 (0%)*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
							GO:0009505, GO:0005886, GO:0009506, GO:0005773, GO:0050660, GO:0008762, GO:0006979,		
116	scaffold98_Gene75_749340_749810	Full=Reticuline oxidase-like protein; Flags: Precursor Length=570	71/155 (46%)	1e-37	Q9SVG4.2	GI:118585329	GO:0031225, GO:0048046, GO:0005829, GO:0005739, GO:0009505, GO:0005886, GO:0009506, GO:0005773, GO:0050660, GO:0008762, GO:0006979,	No hit against flax ESTs	
117	scaffold98_Gene76_750256_751770	Full=Reticuline oxidase-like protein; Flags: Precursor Length=570	62/92 (67%),	5e-34	Q9SVG4.2	GI:118585329	GO:0031225, GO:0048046, GO:0005829, GO:0005739, GO:0009505, GO:0005886, GO:0009506, GO:0005773, GO:0050660, GO:0008762, GO:0006979	*gb|JG284709.1|JG284709 LUSTE1NG-RP-166_A10_10FEB2007_080 LUSTE1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 747, Score = 163 bits (82), Expect = 3e-038, Identities = 142/162 (87%)*	
118	scaffold98_Gene77_751882_752268	Full=Reticuline oxidase-like protein; Flags: Precursor Length=570	58/122 (48%)	2e-32	Q9SVG4.2	GI:118585329	GO:0031225, GO:0048046, GO:0005829, GO:0005739, GO:0009505, GO:0005886, GO:0009506, GO:0005773, GO:0050660, GO:0008762, GO:0006979,	*genolin_c40702 227 nt, Length = 1428, Expect = 8e-005, Identities = 37/41 (90%)*	
119	scaffold98_Gene78_752452_753635	No BLASTx hit against UniProt						*genolin_c21859 375 nt, Length = 1420, Expect = e-161, Identities = 300/304 (98%)*	

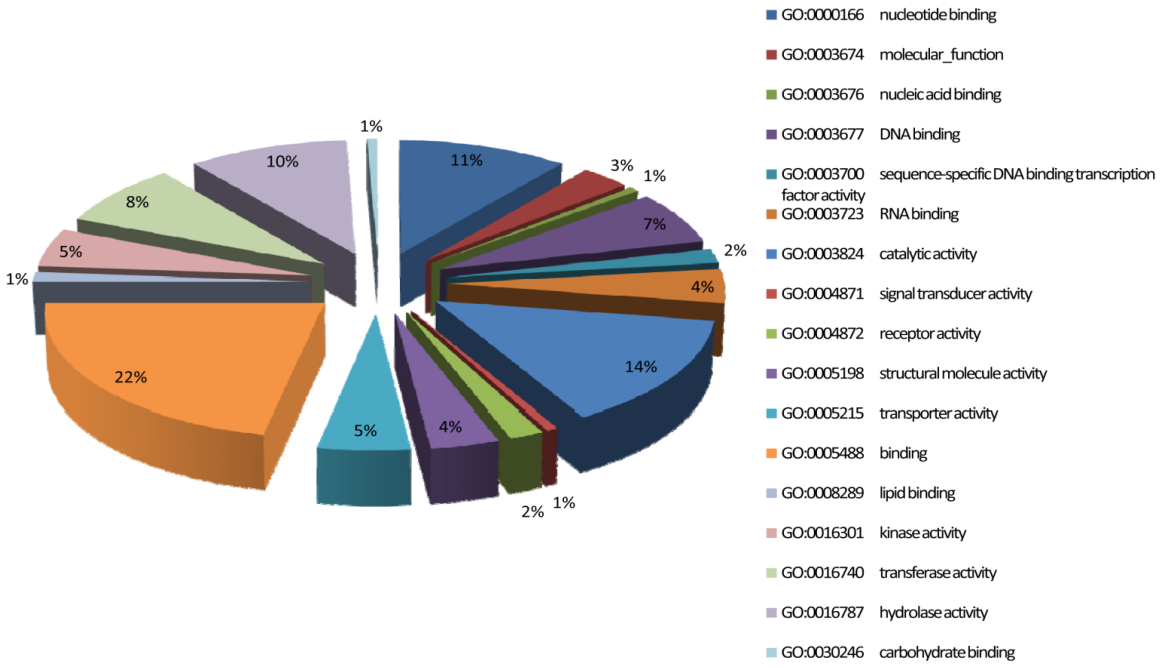
Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
120	scaffold98_Gene79_ 756010 756804 + . ID=Lus10023371; (+ strand)	Full=Intracellular ribonuclease LX; Short=RNase LX; Flags: Precursor Length=237	54/217 (25%)	2e-11	P80196.2	GI:1710616	GO:0005737, GO:0033897, GO:0003723, GO:0090305, GO:0006950,	No hit against flax ESTs	
121	scaffold98_Gene80_ 757680_759425	Full=Putative F-box protein At3g16210 Length=360	52/179 (29%)	4e-10	Q9LU24.1	GI:75274170	No ontology	*gb|EB711981.1|EB711981 LuP12012F11R LuP12 Linum usitatissimum cDNA clone LuP12012F11, mRNA sequence, Length = 407, Expect = 0.0, Identities = 397/407 (97%)*	
122	scaffold98_Gene81_ 759609_760532	Full=Reticuline oxidase-like protein; Flags: Precursor Length=570	118/289 (41%)	2e-72	Q9SVG4.2	GI:118585329	GO:0031225, GO:0048046, GO:0005829, GO:0005739, GO:0009505, GO:0005886, GO:0009506, GO:0005773, GO:0050660, GO:0008762, GO:0006979	*genolin_c40702 227 nt, Length = 1428, Expect = 2e-004, Identities = 37/41 (90%)*	
123	scaffold98_Gene82_ 762275_763798	Full=Reticuline oxidase-like protein; Flags: Precursor Length=570	224/507 (44%	3e-131	Q9SVG4.2	GI:118585329	GO:0031225, GO:0048046, GO:0005829, GO:0005739, GO:0009505, GO:0005886, GO:0009506, GO:0005773, GO:0050660, GO:0008762, GO:0006979,	*gb|JG242658.1|JG242658 LUSTC1NG-RP-143_F11_11MAR2007_085 LUSTC1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 548, Expect = 3e-072, Identities = 319/379 (84%)*	
124	scaffold98_Gene83_ 769673_771286	Full=Reticuline oxidase-like protein; Flags: Precursor Length=570	271/511 (53%)	1e-154	Q9SVG4.2	GI:118585329	GO:0031225, GO:0048046, GO:0005829, GO:0005739, GO:0009505, GO:0005886, GO:0009506, GO:0005773, GO:0050660, GO:0008762,	*genolin_c35272 350 nt, Length = 773, Expect = 0.0, Identities = 363/363 (100%)*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
							GO:0006979,		
125	scaffold98_Gene84_774850_775857	Full=Reticuline oxidase-like protein; Flags: Precursor Length=570	118/355 (33%)	3e-48	Q9SVG4.2	GI:118585329	GO:0031225, GO:0048046, GO:0005829, GO:0005739, GO:0009505, GO:0005886, GO:0009506, GO:0005773, GO:0050660, GO:0008762, GO:0006979,	*genolin_c40702 227 nt, Length = 1428, Expect = 1e-011, Identities = 82/97 (84%)*	
126	scaffold98_Gene85_776136_776447	RecName: Full=Patellin-4 Length=540	60/99 (61%)	2e-35	Q94C59.2	GI:78099068	GO:0005829, GO:0016021, GO:0005634, GO:0005886, GO:0008289, GO:0005215, GO:0007049, GO:0051301,	*LUSTC1NG_RP_011_G02_26FEB2007_004.ab1, Length = 681, Expect = e-174, Identities = 311/312 (99%)*	
127	scaffold98_Gene86_776618_777622	Full=Patellin-4 Length=540	158/335 (47%)	4e-76	Q94C59.2	GI:78099068	GO:0005829, GO:0016021, GO:0005634, GO:0005886, GO:0008289, GO:0005215, GO:0007049, GO:0051301,	*genolin_c28034 619 nt, Length = 1092, Expect = 0.0, Identities = 428/450 (95%), Gaps = 12/450 (2%)*	
128	scaffold98_Gene87_780614_786266	No BLASTx hit against UniProt						*genolin_c32131 396 nt, Length = 644, Expect = 2e-095, Identities = 180/181 (99%)*	
129	scaffold98_Gene88_787152_789714	No BLASTx hit against UniProt						*genolin_c24422 395 nt, Length = 3305, Expect = 2e-077, Identities = 160/164 (97%)*	
130	scaffold98_Gene89_806547_811420	Full=WD40 repeat-containing protein SMU1; AltName: Full=Smu-1 suppressor of mec-8 and unc-52 protein homolog Length=513	53/111 (48%),	3e-20	Q7ZVA0.1	GI:82241387	GO:0005737, GO:0005634,	*genolin_c15411 326 nt, Length = 3176, Expect = 1e-098, Identities = 204/211 (96%)*	
131	scaffold98_Gene90_	No BLASTx hit against						*gb|JG192367.1|JG192367*	

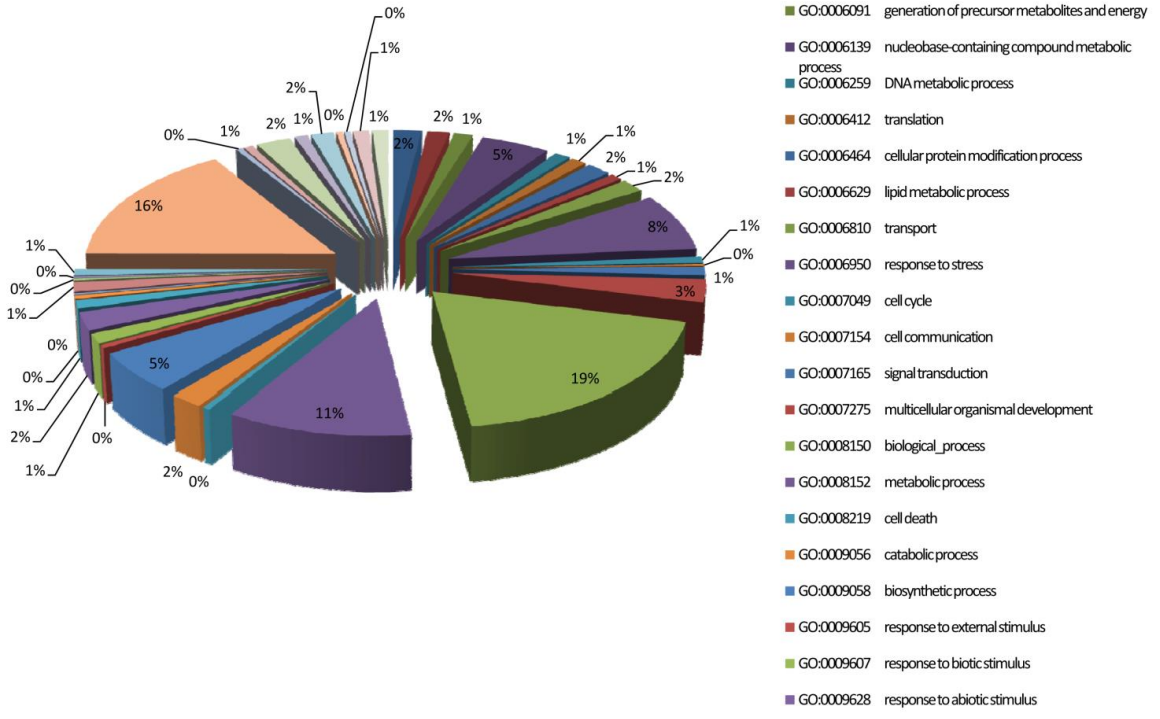
Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
	811430_811661	UniProt						*LUSME1NG-RP-035_D09_21FEB2007_073 LUSME1NG Linum usitatissimum, cDNA, mRNA sequence, Length = 647, Expect = e-128, Identities = 232/232 (100%)*	
132	scaffold98_Gene91_816596_818534	No BLASTx hit against UniProt						*genolin_c25312 373 nt, Length = 372, Expect = 1e-081, Identities = 155/155 (100%)*	
133	scaffold98_Gene92_823055_827063	Full=Pentatricopeptide repeat-containing protein At1g30610, chloroplastic; AltName: Full=Protein EMBRYO DEFECTIVE 2279; Flags: Precursor Length=1006	179/413 (43%),	3e-136	Q9SA76.1	GI:75200328	GO:0009507, GO:0009793,	*gb|EB712625.1|EB712625 LuP12026D03R LuP12 Linum usitatissimum cDNA clone LuP12026D03, mRNA, sequence Length = 399, Expect = e-129, Identities = 238/239 (99%)*	
134	scaffold98_Gene93_837426_839430	Full=Coatomer subunit epsilon-1; AltName: Full=Epsilon-coat protein 1; Short=Epsilon-COP 1; AltName: Full=Epsilon1-COP Length=287	83/127 (65%)	4e-37	Q9MAX6.1	GI:75336169	GO:0030126, GO:0005198, GO:0015031, GO:0006890,	*gb|JG214482.1|JG214482 LUSPS1AD_RP_106_A17_14AUG2008_079 LUSPS1AD Linum usitatissimum cDNA, mRNA sequence, Length = 931, Expect = 3e-088, Identities = 166/166 (100%)*	
135	scaffold98_Gene94_839858_840595	Full=Small nuclear ribonucleoprotein-associated protein B; Short=snRNP-B; AltName: Full=Sm protein B; Short=Sm-B; Short=SmB Length=199	56/91 (62%)	4e-32	Q05856.1	GI:10720262	GO:0015030, GO:0071013, GO:0045495, GO:0071011, GO:0030532, GO:0003723, GO:0007281, GO:0008406, GO:0007052, GO:0000398,	*gb|JG223194.1|JG223194 LUSST4AD-T3-021_O14_15SEP2009_049 LUSST1AD Linum usitatissimum, cDNA, mRNA sequence Length = 927, Expect = 0.0, Identities = 639/642 (99%), Gaps = 1/642 (0%)*	
136	scaffold98_Gene95_841201_847125	Full=60S ribosomal protein L19-3 Length=208	99/105 (94%)	1e-45	P49693.3	GI:19924280	GO:0022625, GO:0005886, GO:0003735, GO:0006412,	*LUSEN1NG_RP_185_D12_28MAR2007_090.ab1, Length = 717, Expect = 0.0, Identities = 472/495 (95%), Gaps = 2/495 (0%)*	

Si. No	Query	BLASTx-Hit (Against UniProtKB database)	alignment length	Identity %	UniPROT KB ID	GI number	GO Id	Blastn hit against Flax-ESTs	Reference
137	scaffold98_Gene96_ 849561_850950	Full=60S ribosomal protein L19-3 Length=208	99/105 (94%)	3e-48	P49693.3	GI:19924280	GO:0022625, GO:0005886, GO:0003735, GO:0006412	*gb|JG217339.1|JG217339 LUSPS1AD_RP_115_C24 _18AUG2008_094 LUSPS1AD Linum usitatissimum, cDNA, mRNA sequence Length = 804, Expect = e-178, Identities = 316/316 (100%)*	
138	scaffold98_Gene97_ 854177_858345	Full=Subtilisin-like protease; AltName: Full=Cucumisin-like serine protease; Flags: Precursor Length=757	50/139 (36%),	2e-16	O65351.1	GI:75099392	GO:0048046, GO:0009505, GO:0004252, GO:0080001, GO:0048359, GO:0043086, GO:0006508,	*gb|JG181098.1|JG181098 LUSLE4AD-T3-051_H19_13NOV2009_07 4 LUSLE1AD Linum usitatissimum cDNA, mRNA sequence Length = 687, Expect = 0.0, Identities = 441/453 (97%)*	
139	scaffold98_Gene98_ 866879_867655	No BLASTx hit against UniProt						No hit against flax ESTs	

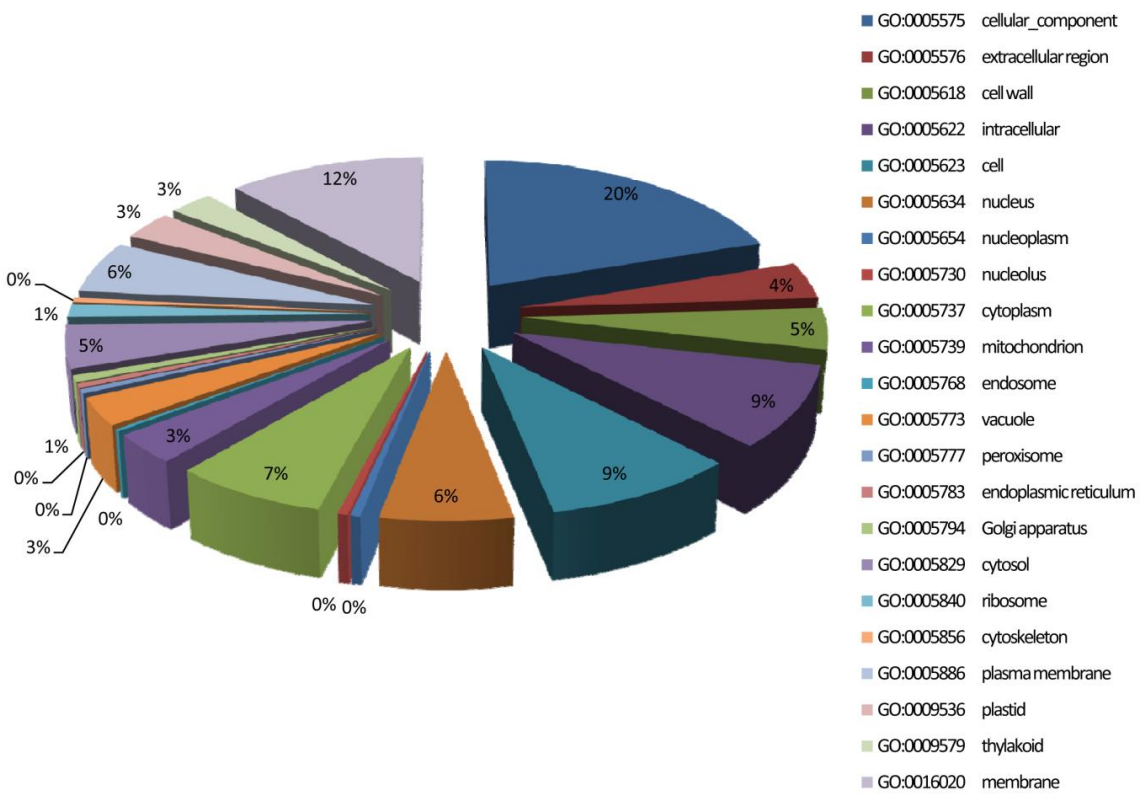
a


1%
11%
3%
1%
7%
2%
4%
14%
2%
1%
4%
5%
22%
1%
5%
8%
10%
GO:0000166 nucleotide binding
GO:0003674 molecular_function
GO:0003676 nucleic acid binding
GO:0003677 DNA binding
GO:0003700 sequence-specific DNA binding transcription factor activity
GO:0003723 RNA binding
GO:0003824 catalytic activity
GO:0004871 signal transducer activity
GO:0004872 receptor activity
GO:0005198 structural molecule activity
GO:0005215 transporter activity
GO:0005488 binding
GO:0008289 lipid binding
GO:0016301 kinase activity
GO:0016740 transferase activity
GO:0016787 hydrolase activity
GO:0030246 carbohydrate binding

b


0%
1%
2%
1%
0%
1%
2%
1%
0%
1%
2%
1%
5%
1%
1%
2%
1%
2%
16%
8%
1%
0%
1%
3%
1%
0%
0%
1%
0%
0%
5%
19%
1%
2%
0%
11%
1%
2%
0%
GO:0000003 reproduction
GO:0005975 carbohydrate metabolic process
GO:0006091 generation of precursor metabolites and energy
GO:0006139 nucleobase-containing compound metabolic process
GO:0006259 DNA metabolic process
GO:0006412 translation
GO:0006464 cellular protein modification process
GO:0006629 lipid metabolic process
GO:0006810 transport
GO:0006950 response to stress
GO:0007049 cell cycle
GO:0007154 cell communication
GO:0007165 signal transduction
GO:0007275 multicellular organismal development
GO:0008150 biological_process
GO:0008152 metabolic process
GO:0008219 cell death
GO:0009056 catabolic process
GO:0009058 biosynthetic process
GO:0009605 response to external stimulus
GO:0009607 response to biotic stimulus
GO:0009628 response to abiotic stimulus

c



Appendix IX GO-slim annotations of gene products predicted from nine non-neutral candidate genomic regions between fiber flax and linseed groups. **a** Molecular function. **b** Biological process. **c** Cellular component.

Appendix X List of Canadian cultivars that are part of the flax core collection.

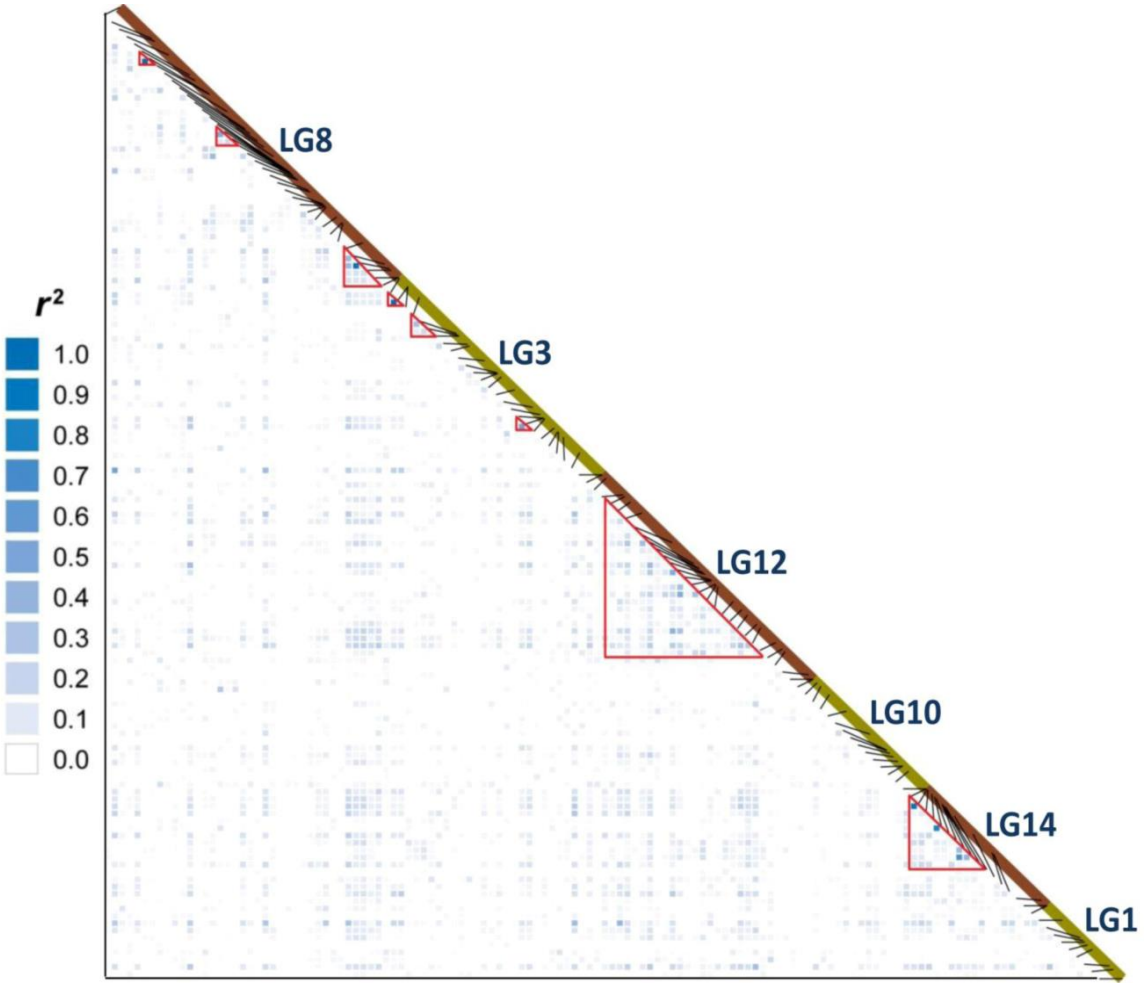
Canadian number	Name	Donor Institute
CN18973	AC Watson	AAFC-Indian Head
CN18979	Flanders	CDC-Saskatoon
CN18980	Somme	CDC-Saskatoon
CN18981	CDC Valour	CDC-Saskatoon
CN19003	AC McDuff	AAFC-Morden
CN19004	AC Emerson	AAFC-Morden
CN19005	AC Linora	AAFC-Indian Head
CN19017	CDC Normandy	CDC-Saskatoon
CN33385	Linott	AAFC-Morden
CN33386	Noralta	AAFC-Morden
CN33388	Redwood 65	AAFC-Morden
CN33389	Rocket	AAFC-Morden
CN33397	Dufferin	AAFC-Morden
CN37286	McGregor	AAFC-Morden
CN52732	Norlin	FP& I Branch, Seed Division
CN100547	Redwing	AAFC-Regina
CN101413	Vimy	CDC-Saskatoon
Linola989	Linola989	
CDCGold	CDCGold	
Macbeth	Macbeth	
Shape	Shape	
CDCSorrel	CDCSorrel	
CDCBethune	CDCBethune	
CDCMons	CDCMons	
CrepitamTabor	CrepitamTabor	
Hanley	Hanley	
Lirina	Lirina	
PrairieBlue	PrairieBlue	
PrairieGrande	PrairieGrande	
PrairieThunder	PrairieThunder	

AAFC: Agriculture and Agri-Food Canada; CDC: Crop Development Center; FP & I Branch: Food Protection and Inspection

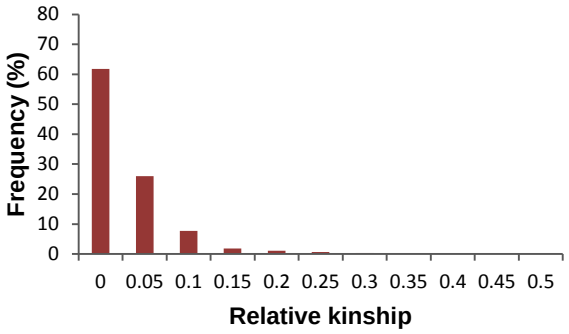
Appendix XI Analysis of variance for seed quality traits in the flax core collection evaluated in six environments. Mean square values and percentage of the total sum of squares for oil content (OIL), palmitic acid (PAL), stearic acid (STE), oleic acid (OLE), linoleic acid (LIO), linolenic acid (LIN), and iodine value (IOD) are shown.

Source of variation	OIL	%TSS[a]	PAL	%TSS[a]	STE	%TSS[a]	OLE	%TSS[a]	LIO	%TSS[a]	LIN	%TSS[a]	IOD	%TSS[a]
Genotype (G)	28.26*	53.8	2.48*	76.9	7.16*	72.1	62.31*	33.3	116.35*	90.6	148.14*	55.1	342.61*	39.0
Location (L)	1166.98*	5.7	28.48*	2.3	259.08*	6.7	19285.56*	26.5	599.85*	1.2	18459.74*	17.6	75216.33*	22.0
Year (Y)	748.63*	7.3	14.65*	2.3	66.19*	3.4	2580.76*	7.1	134.25*	0.5	1830.95*	3.5	6879.74*	4.0
G * L	2.44*	4.6	0.12*	3.7	0.26*	2.6	3.02*	1.6	0.68*	0.5	3.79*	1.4	31392*	3.6
G * Y	2.16*	8.1	0.08*	5.2	0.16*	3.2	2.94*	3.1	1.04*	1.6	3.52*	2.6	28.13*	6.4
L * Y	354.34*	3.5	1.27*	0.2	62.97*	3.3	3376.52*	9.3	160.11*	0.6	2904.21*	5.6	11748.27*	6.9
G * L * Y	1.73*	5.7	0.07 *n.s.*	4.1	0.13*	2.4	2.32*	2.3	0.53*	0.8	2.88*	2.0	26.61*	5.7

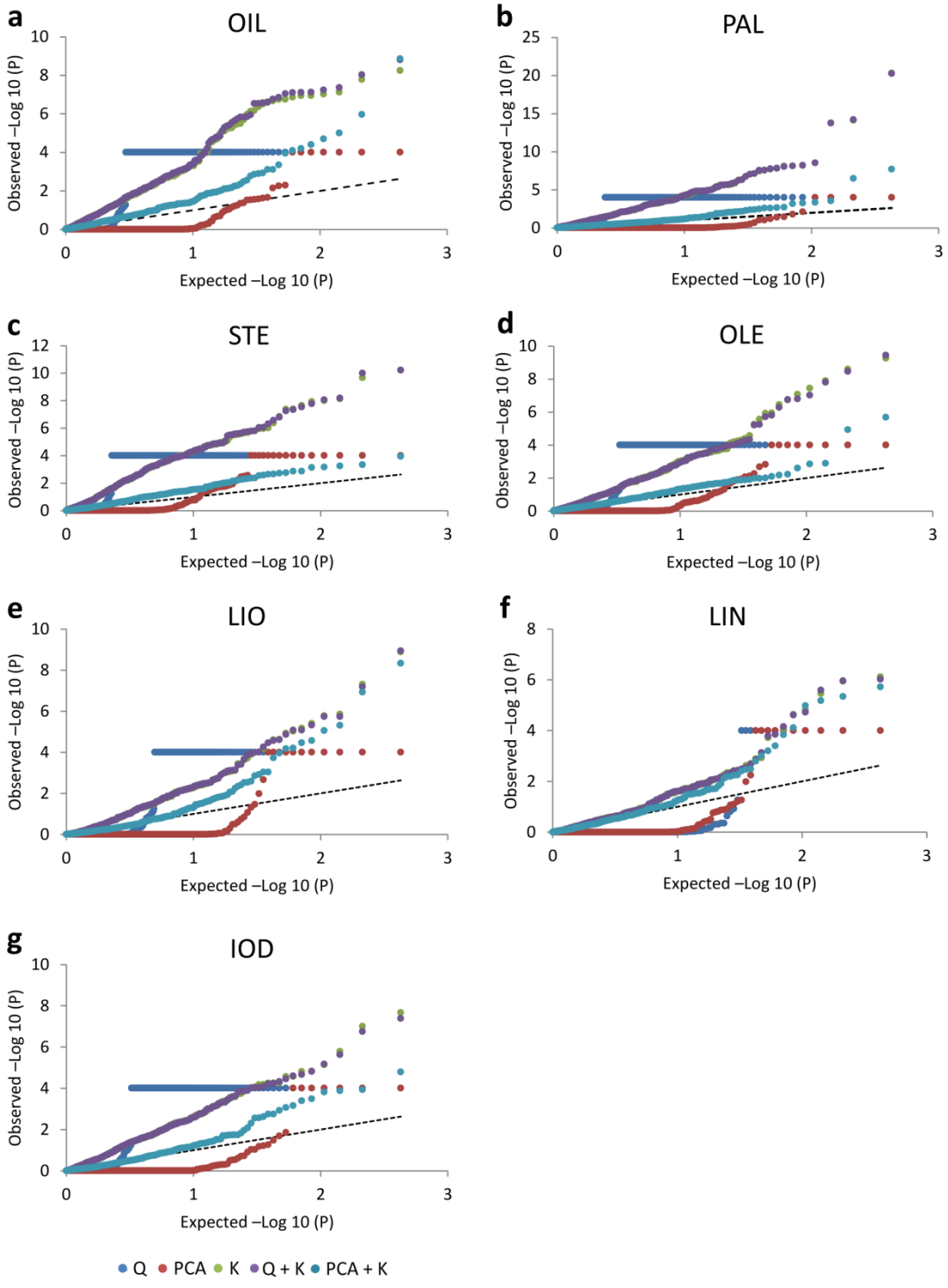
[a] % TSS = percentage of the total sum of squares
* Significant at $P < 0.0001$; *n.s.* = non-significant



Appendix XII Linkage disequilibrium (LD) heat map of six linkage groups (LGs) in linseed. Red triangles highlight blocks of LD across LGs and the colored ruler indicates the strength of LD (r^2).



Appendix XIII Pairwise relative kinship estimates of the flax core collection based on 448 microsatellite markers.



Appendix XIV Cumulative probability-probability (P-P) plots of the observed -Log_{10} (P) values (y-axes) against the expected distribution (dotted diagonal line) of -Log_{10} (P) values (x-axes) for the general linear model (*Q*), the general linear model (PCA), the mixed linear model (*K*),the mixed linear model (*Q* + *K*) and the mixed linear model (PCA + *K*). **a** oil content (OIL) **b** palmitic acid content (PAL) **c** stearic acid content (STE) **d** oleic acid content (OLE) **e** linoleic acid content (LIO) **f** linolenic acid content (LIN) **g** iodine value (IOD).

Appendix XV Candidate QTL associated with seven seed quality traits identified at either or both of the Manitoba (MB) and Saskatchewan (SK) locations.

Trait	Contig-Scaffold-Marker	LG	Position	LOC	-Log_{10} (P)	R^2 (%)	Effect[a]	Favorable allele (bp)
OIL	c108-s305_Lu2649	8	18.92	MB	4.19	1.72	0.66*	228
	c31-s67_Lu181	9	31.34	MB	3.47	3.70	1.12*	270
				SK	3.65	8.31	1.40**	270
	c31-s8_Lu2262	9	23.58	SK	3.50	1.77	1.69**	186
	c77-s151_Lu519	9	32.54	MB	3.30	1.49	0.81*	239
				SK	3.32	4.61	0.96**	239
	c0-s0_Lu926Bb	12	32.87	MB	3.81	2.03	0.62*	596
	Lu242	-	-	MB	3.30	2.83	1.53**	324
	Lu401	-	-	MB	4.11	9.66	1.70**	317
	Lu788	-	-	MB	3.41	1.07	0.92**	280
PAL	c7-s471_Lu2040	3	74.36	MB	6.83	1.41	0.30*	147
				SK	8.05	3.32	0.31*	147
	c79-s540_Lu2534	7	5.79	MB	6.67	1.01	0.90**	312
				SK	6.60	2.09	0.71**	312
	c38-s34_Lu2046	11	0.00	MB	3.10	8.44	1.01**	152
	c214-s863_Lu2917a	12	0.00	MB	3.29	4.24	0.35**	215
	Lu681	-	-	SK	4.37	1.27	0.33*	310
STE	c222-s821_Lu943	1	149.99	MB	3.41	2.04	0.62**	274
	c160-s260_Lu747b	2	0.00	SK	3.40	1.03	0.96**	261
	c400-s216_Lu3150	3	113.37	MB	3.62	8.34	1.17**	366
	c118-s196_Lu558	3	122.39	MB	4.06	1.26	0.21*	265

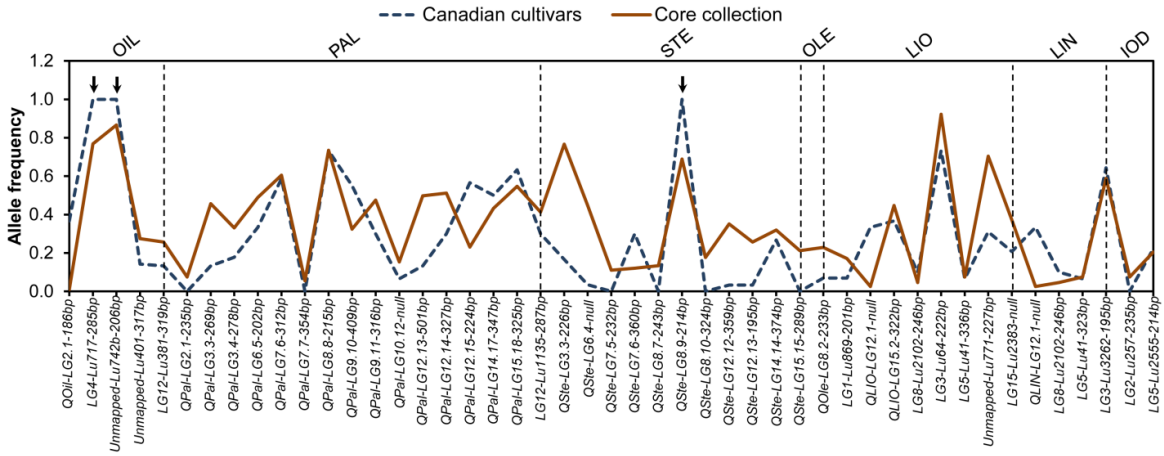
Trait	Contig-Scaffold-Marker	LG	Position	LOC	-Log_{10} (P)	R^2 (%)	Effect[a]	Favorable allele (bp)
				SK	3.37	5.91	0.24*	265
	c171-s62_Lu1112	6	84.84	MB	3.40	9.18	0.39*	270
	c79-s511_Lu2532	7	2.73	MB	3.64	4.92	1.64**	270
				SK	3.51	3.34	1.29**	270
	c175-s1216_Lu146	7	23.95	MD	3.32	13.2	1.01**	354
		7	23.95	SK	3.37	19.68	1.00**	354
	c175-s1216_Lu151	7	23.99	SK	3.20	9.07	0.73**	286
	c0-s635_Lu928	8	74.30	SK	3.49	1.51	1.01**	243
	c32-s0_Lu2279	13	37.19	MB	3.51	3.63	0.79**	210
	Lu316	-	-	MB	3.42	10.21	0.69**	223
				SK	3.40	3.79	0.49**	223
	Lu319	-	-	MB	3.40	2.13	0.55**	230
	Lu401	-	-	MB	3.63	1.04	0.51**	317
STE	Lu707	-	-	MB	3.54	9.26	0.92**	538
OLE	c82-s1491_Lu2564	6	64.09	MB	3.96	1.18	3.66**	251
				SK	5.41	1.94	2.72**	251
	c82-s176_Lu2555	6	72.00	MB	4.24	3.97	2.45**	217
				SK	4.84	3.93	2.07**	217
LIO	**c729-s156_Lu3262**	3	55.74	MB	5.99	8.34	0.93*	195
				SK	5.18	6.91	0.75*	195
	c0-s156_Lu64	3	60.88	MB	3.76	2.75	3.94**	220
				SK	4.12	2.88	3.62**	220
	c16-s156_Lu373	3	64.44	MB	3.95	1.79	2.94**	216

Trait	Contig-Scaffold-Marker	LG	Position	LOC	-Log_{10} (P)	R^2 (%)	Effect[a]	Favorable allele (bp)
				SK	3.80	1.02	2.73**	216
	c202-s39_Lu41	5	57.36	MB	7.41	0.89	0.65*	336
				SK	6.41	0.82	0.62*	336
	c30-s11_Lu164	5	57.89	MB	3.60	1.72	0.53*	211
				SK	3.50	1.93	0.54*	211
	c436-s86_Lu296	7	40.77	MB	3.60	3.31	2.05**	526
	c436-s86_Lu672	7	43.41	MB	3.82	2.72	2.10**	197
	c108-s159_Lu585B	7	53.67	MB	3.63	1.07	0.99*	208
				SK	3.50	1.11	0.93*	208
	c281-s1851_Lu566	7	91.55	MB	5.11	1.08	1.19**	214
				SK	5.32	1.12	1.08**	214
	c82-s617_Lu2561a	8	28.95	MB	4.19	0.76	0.78**	338
				SK	3.67	0.89	0.70**	338
	c46-s505_Lu2102	8	72.74	MB	7.73	2.96	0.77*	241
				SK	9.65	7.57	0.63*	241
	c306-s98_Lu206b	12	71.90	MB	4.39	1.07	0.56*	null
				SK	3.96	1.41	0.54*	null
	c306-s98_Lu203b	12	72.55	MB	4.03	0.93	0.51*	null
				SK	3.58	1.41	0.58*	null
	c306-s98_Lu765Bb	12	75.12	MB	5.05	4.76	0.97*	null
				SK	4.84	4.42	0.90*	null
	Lu771	-	-	MB	3.60	2.99	1.17*	230
				SK	3.59	1.81	1.46**	230

Trait	Contig-Scaffold-Marker	LG	Position	LOC	-Log_{10} (P)	R^2 (%)	Effect[a]	Favorable allele (bp)
LIN	**c729-s156_Lu3262**	3	55.74	MB	4.17	3.68	1.17*	195
				SK	4.61	5.26	1.34*	195
	c16-s156_Lu373	3	64.44	MB	3.50	1.80	2.10**	216
LIN				SK	4.62	4.85	1.67**	216
	c30-s11_Lu164	5	57.89	MB	3.63	2.65	1.81**	211
				SK	3.51	3.07	1.24**	211
	c202-s39_Lu41	5	57.36	MB	4.22	1.90	2.04**	323
				SK	4.81	3.45	1.54*	323
	c108-s159_Lu585B	7	53.67	MB	3.60	0.98	0.71*	208
				SK	3.51	1.32	0.79*	208
	c281-s1851_Lu566	7	91.55	MB	3.60	0.92	0.64*	214
				SK	3.86	0.96	0.77*	214
	c82-s617_Lu2561a	8	28.95	MB	4.15	1.85	2.26**	338
				SK	4.83	1.78	2.04**	338
	c46-s505_Lu2102	8	72.74	MB	5.32	3.63	0.74*	241
				SK	6.85	5.07	0.78*	241
	c141-s641_Lu2746	10	87.69	MB	3.60	4.31	2.40**	406
				SK	3.50	2.55	1.73**	406
	c306-s98_Lu206b	12	71.90	MB	3.84	0.93	0.76*	214
				SK	4.20	0.92	0.79*	214
	c306-s98_Lu203b	12	72.55	MB	3.60	1.20	0.79*	214
				SK	3.69	1.31	0.72*	214
	c306-s98_Lu765Bb	12	75.12	MB	3.60	1.18	0.92*	null

Trait	Contig-Scaffold-Marker	LG	Position	LOC	-Log_{10} (P)	R^2 (%)	Effect[a]	Favorable allele (bp)
				SK	4.40	2.33	0.87*	null
IOD	c28-s475_Lu2247	2	50.86	MB	3.96	2.21	9.8**	285
	c729-s156_Lu3262	3	55.74	SK	4.43	1.43	4.99**	195
	c82-s176_Lu2555	6	72.00	MB	3.84	4.83	8.58**	214
				SK	8.85	5.13	6.78**	214
	c82-s617_Lu2561a	8	28.95	MB	3.61	2.12	5.90**	338
	c46-s505_Lu2105	8	72.13	MB	3.60	2.28	13.60**	226
	c46-s505_Lu2102	8	72.74	SK	4.23	9.35	9.31**	241
	c0-s635_Lu928	8	74.30	MB	3.89	3.03	9.66**	247
				SK	3.49	3.53	9.30**	247
	c141-s641_Lu2746	10	87.69	MB	3.60	1.42	5.34**	406
	Lu720	-	-	SK	3.41	0.89	7.76**	321

[a]Effect of favorable alleles represents the increment in percentage of FAs. For IOD the effect represent iodine value in units
Significance of the allelic effects tested by Kruskal-Wallis non-parametric test *$P < 0.01$; **$P < 0.001$
Markers in bold script represent the candidate QTL

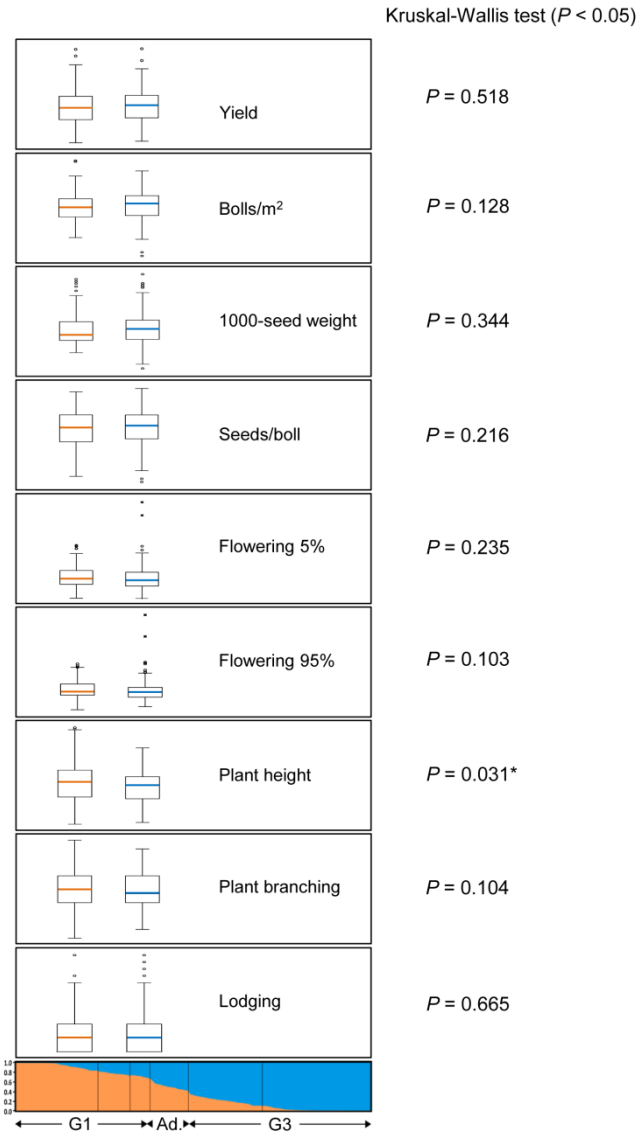


Appendix XVI Comparison of the frequency of favourable QTL/marker alleles across seven quality traits in 30 linseed Canadian cultivars and the remaining 377 accessions of the flax core collection. Arrows indicate fixed alleles in Canadian germplasm.

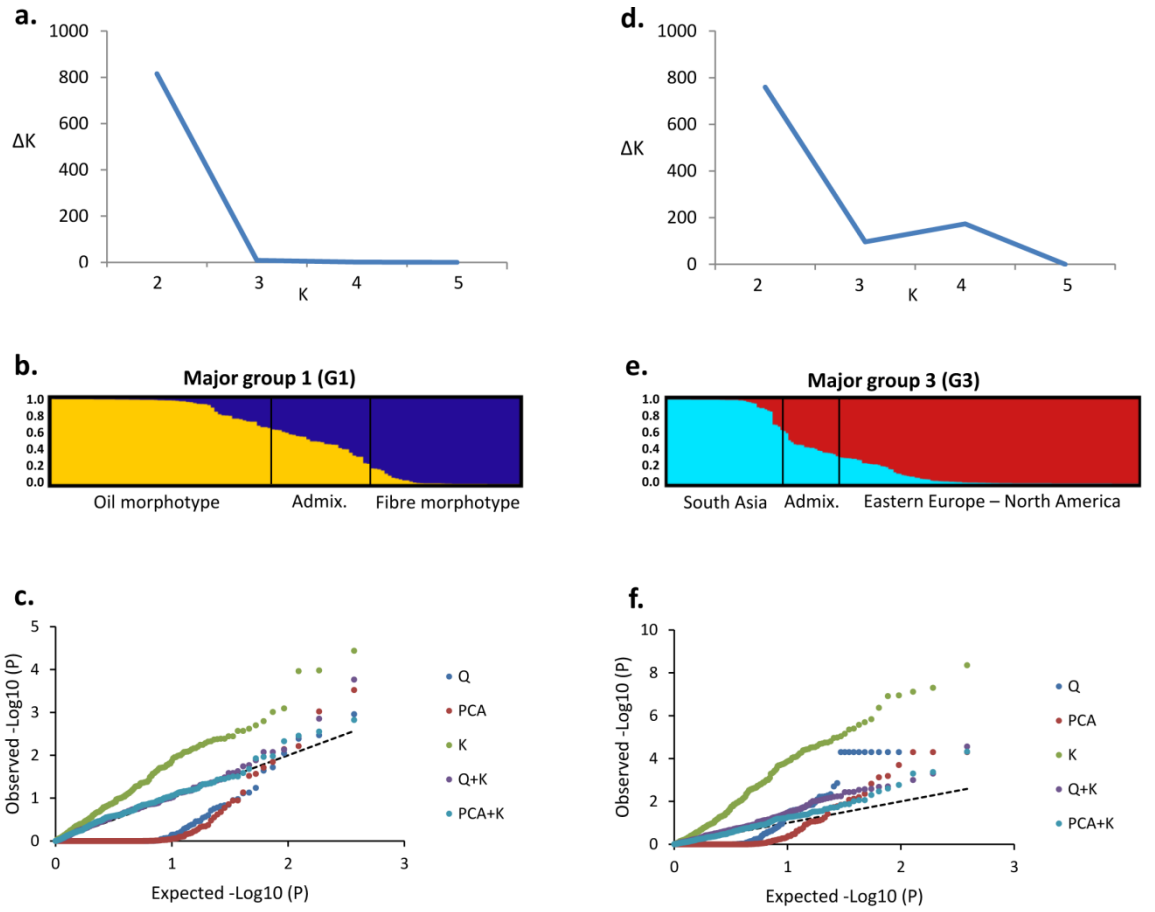
Appendix XVII Analysis of variance for nine agronomic traits in the flax core collection. Yield, bolls per area (BPA), 1000-seed weight (TSW), seeds per boll (SPB), start of flowering (FL5%), end of flowering (FL95%), plant height (PH), plant branching (PBR) and lodging (LDG).

	Source of variation							
Trait	Environments	Genotype (G)	Location (L)	Year (Y)	G*L	G*Y	L*Y	G*L*Y
Yield	6	1165352.55**	30565660.13**	51027784.21**	219834.27*n.s.*	292173.54*n.s.*	6640745.72**	280249.62*n.s.*
BPA	6	3798562.85**	155515107.49**	103746480.06**	1297049.31*	1611240.63**	17454489.68**	1117489.13*n.s.*
TSW	6	5.31**	12.61**	323.13**	0.51**	0.81**	76.46**	0.63**
SPB	6	5.54**	1315.95**	8.07**	1.10**	1.50**	47.67**	0.93**
FL5%	7	47.17**	12747.33**	4401.04**	14.14**	8.87**	102.93**	10.76**
FL95%	7	44.84**	10372.19**	4060.71**	11.33**	8.57**	1238.15**	11.31**
PH	6	707.53**	218231.43**	56488.44**	94.71**	175.65**	17283.19**	65.53**
PBR	4	1.66**	50.89**	6.11**	1.82**	0.96**	-	-
LDG	8	1.40**	58.62**	65.19**	0.72**	0.76**	40.75**	0.69**

* and ** significant at $P < 0.01$ and $P < 0.001$, respectively; *n.s.* = non-significant



Appendix XVIII Box plots for nine agronomic traits for the two major groups, G1 and G3. Significant differences for traits between major groups were tested by Kruskal-Wallis test ($P < 0.05$).



Appendix XIX Structure analysis within major groups and model comparison for plant height. **a** and **d** estimation of the most probable number of subgroups (*K*) using the *ad-hoc* Δ*K* (Evanno et al. 2005) for *K* values ranging from 1 to 5 within G1 and G3, respectively **b** and **e** estimation of the hypothetical number of subpopulations using STRUCTURE (Pritchard et al. 2000) within G1 and G3, respectively. Each individual is represented by a vertical column partitioned into *K* colored segments proportional to their coefficient of membership (*Q*) to each subpopulation **c** and **f** probability plots (P-P) of observed versus expected -Log_{10} (P) values for PH evaluated with five association mapping models within G1 and G3, respectively. *Q* general linear model using the *Q* matrix, PCA general linear model using the PCA matrix, *K* mixed linear model using the kinship matrix, *Q* + *K* mixed linear model using the *Q* and PCA matrices, PCA + *K* mixed linear model using the PCA and *K* matrices.